


RESEARCH ARTICLE OPEN ACCESS

Regional Shopping Objectives in British Grocery Retail Transactions Using Segmented Topic Models

Mariflor Vega Carrasco¹ | Mirco Musolesi² | Jason O'Sullivan³ | Rosie Prior³ | Ioanna Manolopoulou⁴ 

¹Microsoft, London, UK | ²Department of Computer Science, University College London, London, UK | ³Dunnhumby Ltd., London, UK | ⁴Department of Statistical Science, University College London, London, UK

Correspondence: Ioanna Manolopoulou (i.manolopoulou@ucl.ac.uk)

Received: 12 April 2024 | **Revised:** 6 August 2024 | **Accepted:** 12 August 2024

Funding: This work was supported by ESRC Grant No. ESRC ES/P000592/1 and The Alan Turing Institute under the UK EPSRC Grant No. EP/N510129/1.

Keywords: latent Dirichlet allocation | retail analytics | segmented topic model | spatial modelling | topic modelling

ABSTRACT

Understanding the customer behaviours behind transactional data has high commercial value in the grocery retail industry. Customers generate millions of transactions every day, choosing and buying products to satisfy specific shopping needs. Product availability may vary geographically due to local demand and local supply, thus driving the importance of analysing transactions within their corresponding store and regional context. Topic models provide a powerful tool in the analysis of transactional data, identifying topics that display frequently-bought-together products and summarising transactions as mixtures of topics. We use the segmented topic model (STM) to capture customer behaviours that are nested within stores. STM not only provides topics and transaction summaries but also topical summaries at the store level that can be used to identify regional topics. We summarise the posterior distribution of STM by post-processing multiple posterior samples and selecting semantic modes represented as recurrent topics, and employ Gaussian process regression to model topic prevalence across British territory while accounting for spatial autocorrelation. We implement our methods on a dataset of transactional data from a major UK grocery retailer and demonstrate that shopping behaviours may vary regionally and nearby stores tend to exhibit similar regional demand.

1 | Introduction

In the grocery retail industry, millions of transactions are generated every day by customers that choose and buy products to fulfil one or more needs. Transactions typically contain few products out of thousands of available items, reflecting unseen customer motivations. Customers visit grocery retailers to fulfill different shopping objectives; for instance, to buy food for breakfast, ingredients to cook a roast dinner or popular products for a barbecue. Identifying customer behaviours provides insights into high-resolution shopping patterns that may help retailers to maximise efficiency while delivering value to all stakeholders.

Customer motivations may be driven not only by everyday needs, but also by geographical effects, that is, showing product combinations that are only relevant in specific regions. For example, a store in Scotland may offer products from local brands and/or products that are part of the local cuisine; these products might not have the same popularity in other further constituent countries in the UK. In response, retailers customise product assortments in each store to include locally supplied products and to fulfil local demand. Identifying regional shopping objectives may help retailers launch marketing campaigns, customise store assortments and layout, and may also support the prediction of composition for new stores. Moreover, geographical

Abbreviations: GP, Gaussian process; STM, segmented topic model.

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2024 The Author(s). *Applied Stochastic Models in Business and Industry* published by John Wiley & Sons Ltd.

resolution of shopping patterns may aid the investigation of eating habits driven by social and cultural factors that otherwise rely on expensive ad hoc studies.

In this article, we aim to model the regional distribution of shopping objectives represented as grocery product combinations directly from transactions. We analyse transactions from a major grocery retailer in the UK. The data consist of individual transactions, indexed by store, where each observation is a set of products purchased within a single transaction at a particular store within a particular timeframe. Our overarching goal is to identify combinations of products in high demand in specific areas and to determine their spatial prevalence, characterising customer behaviours from different regions and constituent countries of the UK. To this end, two main ingredients are needed: a model for capturing customer behaviours through transactional data, combined with a model that captures the spatial distribution of these customer behaviours.

To model grocery transactions, we employ the segmented topic model (STM) [1], which models product combinations purchased together as a mixture of multinomial distributions representing ‘shopping objectives’, taking into account the store hierarchy. This enables the identification of product combinations within the store context. STM provides topic distributions (here representing shopping objectives), transaction-specific mixtures of objectives, and store-specific mixtures of objectives. Shopping objectives describe products that are frequently bought together with high probabilities, reflecting different customer behaviours. Mixtures of objectives summarise purchased products according to their composition of objectives, that is, a very popular shopping objective in a store-specific mixture of objectives would show a high probability.

To complement the transaction model with a spatial distribution without a fully integrated model, we extract posterior summaries for each shopping aim of the STM to feed into a Gaussian process regression (GPR) [2]. This allows us to capture multiple posterior modes of the STM, so that uncertainty about shopping objectives can be meaningfully incorporated through posterior clusters. We characterise the spatial distribution of regional shopping objectives using GPR [3–5] on store-specific objective probabilities, modelling topical prevalence over the UK. GPR accounts for store meta-data and the geographical proximity between stores, both of which STM does not account for, and allows us to identify and characterise variation in the topic probabilities, which are explained by spatial autocorrelation as well as regional covariates. We demonstrate that the GPR approach naturally achieves a better out-of-sample predictive behaviour than a linear model by borrowing information from neighbouring stores while affording an interpretable model with quantifiable uncertainty.

This article is organised as follows: Section 2 provides the contextual background to topic modelling and spatial modelling. STM and GPR are introduced in Sections 3 and 4. Regional topics in British grocery retail transactions are presented in Section 5. Spatial analysis of regional topics is discussed in Section 6. Finally, we conclude and summarise our findings in Section 7.

2 | Background

Topic modelling was originally introduced to analyse and summarise large collections of text corpora. In retail analytics, topic modelling allows describing transactions and groups of transactions as probabilistic mixtures of topics, representing shopping objectives, which are distributions over a fixed product assortment. For the remainder of this article, we will use ‘topics’ and ‘shopping objectives’ interchangeably, depending on the context. Different shopping objectives exhibit different combinations of products with high probability, expressing different customer behaviours. Topic models have been applied to model customer behaviour over highly aggregated product assortments [6–10], but were only recently applied to transactional data at the full product resolution [2, 11].

In the UK, spatial analysis has been previously applied to grocery retail data to study store catchment and store performance. For example, Sturley, Newing, and Heppenstall [12] used an agent-based model to extract key consumer behaviours about shopping frequency, shopping mission, store choice and spending. Davies, Dolega, and Arribas-Bel [13] applied a spatial interaction modelling (SIM) technique, to create catchment areas and investigate the spatial variation on competition, sales area, trade intensity, among other factors. With a SIM approach, Newing, Clarke, and Clarke [14] forecasted store patronage and store revenues in two English regions. Waddington et al. [15] explored spatiotemporal fluctuations of store sales and catchment areas. Berry et al. [16] examined workplace geographies and census statistics to investigate store trading characteristics in inner London. However, none of the existing literature investigates spatial variations of customer behaviours by modelling product combinations directly.

Grocery consumption habits are an important driver of population health, and in the UK are studied regularly by the Department for Environment Food and Rural Affairs (DEFRA) through the targeted, detailed ‘Family Food Survey’ and by the Office for National Statistics through the former ‘Living Costs and Food Survey’ [17]. Other studies use targeted consumer surveys [18] or analyse coarse consumer trends such as the ‘Greggs-Pret’ index [19] to explore regional differences. These studies offer a valuable resource and provide opportunities to explore many aspects of purchasing patterns: expenditure, intake of energy and nutrients, geographical and demographic differences, and trends over time. However, surveys are costly and time-consuming to obtain, as they require participants to keep a food diary over a period of several weeks, which is analysed alongside participants’ answers to interview questions. In contrast, the data used in this article are readily available through stores’ transaction records, but are high in resolution as they capture individual products.

On the other hand, regional food consumption has also been discussed in anthropological and sociological works. For instance, Kuznesof, Tregear, and Moxey [20] found that ‘regional foods’ are perceived as ‘regional products’ or ‘regional recipes’, which are associated with high-value, speciality, or hand-crafted products and with dishes that require home preparation and cooking. Groves [21] defined ‘regional food’ as the food of a particular area

of the country, often representing a regional speciality. However, these studies were also not carried out using transactional data, but instead employed market research methods such as focus groups and questionnaires.

In this article, we capture topics with geographical variability by accounting for the dependency of transactions on store index, to reflect store-specific product assortment, that is, transactions can only contain products that are available at their associated stores. Modelling this dependency implies a hierarchy in which stores are one level above transactions. Without store hierarchy, regionally purchased products would be drowned out by the sheer volume of nationally supplied products, hampering the identification of regional topics. Thus, we apply the STM [1], which enables the identification of product combinations within the store context. STM provides topic distributions, transaction-specific mixtures of objectives, and store-specific mixtures of objectives.

Summarising the posterior distribution of STM is needed but it is not an easy task. Topic models are often highly multi-modal, resulting in topics that may not reappear among posterior samples [22, 23]. Here, we summarise posterior topic distributions by identifying thematic modes following the clustering methodology in [2]. This methodology fuses topic distributions from multiple posterior samples to identify recurrent topics and their associated uncertainties. Topics are grouped into clusters, which are represented by their average distribution, named *clustered* topics, and by their cluster size, named *recurrence*. Users evaluate subsets of clustered topics and select a posterior topical summary depending on generalisation and quality metrics.

2.1 | Connections to Other Models

Fully integrating STM and GPR by assuming a Gaussian Process distribution over store-specific mixtures of objectives is feasible but computationally prohibitive. The joint model would be similar to the correlated topic model [24] with two layers of mixtures of objectives and a covariance matrix defined over geographical distance. We do not pursue this approach as the non-conjugacy of the Gaussian process poses a challenge for posterior inference. Instead, we feed a GPR with a topical posterior summary obtained from the clustering of the STM posterior samples, taking advantage of the closed form Gibbs sampler of STM.

Although in this and related work we have performed extensive model comparison and model exploration, direct model validation (for example, through simulated data) is challenging. This is because our approach does not have an obvious data generative model, since it is fitted stepwise. The closest generative model is perhaps a combination of the STM used in this article [1], extended to include covariate effects on the logit scale of the topic probabilities, similarly to the structural topic model [25]. In this case, the covariate effects would include a fixed effect of the region along with spatially-varying errors through a GP regression. Although fitting this model would be very challenging, one could use such a model to create simulated data for the purpose of model validation for our two-step approach. This, however, is beyond the scope of the current paper and is part of ongoing work.

3 | Topic Modelling

Topic modelling was originally introduced to automatically organise, understand, and summarise large collections of text corpora. Latent Dirichlet allocation (LDA) [26, 27] is one of the most popular topic modelling techniques, which represents documents as mixtures of topics, and topics as distributions over a fixed vocabulary. The STM [1] is an extension to LDA which includes hierarchical structure within documents, thereby representing documents as collections of paragraphs (segments). Both documents and paragraphs are represented as mixtures of topics, where a paragraph-specific topical mixture derives from its document-specific topical mixture. LDA and STM interpret documents as *bags of words*, disregarding word order.

STM has been mainly used in text applications but has not been applied in retail analytics to the best of our knowledge. For instance, STM has been used to match experts with questions [28] and to analyse multi-aspect sentiment in customer reviews [29]. We apply STM in the context of grocery retail data, interpreting stores as documents, transactions as segments and topics as distributions over a fixed assortment of products. Transaction-specific mixtures of objectives derive from the corresponding store-specific topical mixture. Thus, transactions and stores share the space of latent topics; placing a hierarchical structure over transactions at each store allows us to capture socio-economic and cultural variability across different areas. The *bag of words* assumption organically fits the grocery retail domain since products are registered at stores without an inherent order.

In the standard LDA model, topics display products that are frequently purchased together. If a product is frequently purchased in few stores (and rarely purchased due to unavailability or low preference in the majority of stores), then the product is unlikely to rank highly within a topic. Thus, analysing retail data through LDA might overlook topics that reflect regional or local customer behaviours. In contrast, STM can harness meta information of store hierarchy over transactions. Thereby, product co-occurrence is relative to store context and transactions taking place at the same store are expected to exhibit more similar mixtures of objectives than transactions from other stores.

3.1 | Segmented Topic Model

STM [1] consider the following hidden variables: topic distributions, store-specific mixtures of objectives and transaction-specific mixtures of objectives. In detail, K topic distributions, $[\phi_1, \dots, \phi_K]$, are sampled from a Dirichlet distribution governed by hyperparameters β ; each ϕ is a V -dimensional vector, and V is the size of the product assortment. D store-specific mixtures of objectives, $\theta_1, \dots, \theta_D$, are sampled from a Dirichlet distribution governed by hyperparameters α , where D is the number of stores; each θ is a K -dimensional vector. P transaction-specific mixtures of objectives, $\nu_{1,d}, \dots, \nu_{P,d}$, are sampled from a Poisson–Dirichlet process (PDP) distributed with discount parameter a , strength parameter b and base measure θ_d ; each ν is also a K -dimensional vector.

STM follows a generative process in which each transaction is created by sampling products from topics, which are also

sample from a transaction-specific topical mixture. This generative process has two steps. First, a topic assignment $z_{n,p,d}$ is sampled from a transaction-specific topical mixture $\nu_{p,d}$. Second, a product $w_{n,p,d}$ is sampled from the assigned topic distribution $\phi_{z_{n,p,d}}$, where n is the n th item in transaction p in store d . Mathematically,

$$\begin{aligned}
 \phi_k &\sim \text{Dirichlet}(\boldsymbol{\beta}) \\
 \theta_d &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\
 \nu_{p,d} &\sim \text{PDP}(a, b, \theta_d) \\
 z_{n,p,d} &\sim \text{Multinomial}(\nu_{p,d}) \\
 w_{n,p,d} &\sim \text{Multinomial}(\phi_{z_{n,p,d}})
 \end{aligned} \tag{1}$$

The PDP [30–32] is a generalisation of the Dirichlet process, also called the Pitman–Yor process. PDP is useful to handle conjugacy between Dirichlet and Multinomial distributions. A graphical representation of the STM is shown in Figure 1.

3.1.1 | Inference

PDP has a useful representation through the Chinese restaurant process (CRP) [33]. CRP follows an intuitive analogy in which a Chinese restaurant with infinite *table* capacity receive customers who choose to sit around an occupied table or to open a new table; customers sitting around the same table share the same dish. Interpreting the CRP in the retail context, customers are products and dishes are customer behaviours; thus products that fulfil the same customer need are grouped around the same topic. Note that customers and dishes are linked through tables and a dish can be served by multiple tables. Thus, the CRP introduces ‘table counts’, constrained latent variables \mathbf{t} , that represent the number of tables serving the same dish.

Marginalising transaction-specific variables ν introduces the constrained latent variables \mathbf{t} and leaves the store-specific variables θ in conjugate form. Integrating out topic distributions ϕ and mixtures of objectives θ, ν , the joint conditional distribution of STM is:

$$\begin{aligned}
 p(\mathbf{z}, \mathbf{w}, \mathbf{t} | \boldsymbol{\alpha}, \boldsymbol{\beta}, a, b) &= \prod_d \frac{\text{Beta}_K(\boldsymbol{\alpha} + \sum_p \mathbf{t}_{p,d})}{\text{Beta}_K(\boldsymbol{\alpha})} \prod_{p,d} \frac{(b|a)^{\sum_k t_{p,d,k}}}{(b)^{N_{p,d}}} \\
 &\prod_{p,d,k} S_{t_{p,d,k},a}^{N_{k|p,d}} \prod_k \frac{\text{Beta}_V(\boldsymbol{\beta} + \mathbf{N}_k)}{\text{Beta}_V(\boldsymbol{\beta})} \tag{2}
 \end{aligned}$$

where $t_{p,d,k}$ is the table count for transaction p , store d and topic k . $\text{Beta}_K(\boldsymbol{\alpha})$ is the K -dimensional beta function that normalises the Dirichlet distribution; $\mathbf{t}_{p,d} = [t_{p,d,1}, \dots, t_{p,d,K}]$ is a K -dimensional vector of table count; $(x|y)_N$ denotes the Pochhammer symbol; $N_{p,d}$ size of transaction p in store d ; $S_{M,a}^N$ is a generalised Stirling number; $N_{k|p,d}$ number of topic assignments of topic k in transaction p in store d . $\text{Beta}_V(\boldsymbol{\beta})$ is V dimensional beta function that normalises the Dirichlet distribution; $\mathbf{N}_k = [N_{1|k}, \dots, N_{v|k}, \dots, N_{V|k}]$ is a V -dimensional vector of term counts, which is the number of products of type v assigned to topic k . Detailed definitions of the Pochhammer symbol and generalised Stirling number are explained in [1].

Due to the intractable computation of marginal probabilities, the posterior distribution of latent variables cannot be computed directly. Thus, inference of STM uses a Monte Carlo approximation through a Gibbs sampler. Du, Buntine, and Jin [1] and Buntine and Hutter [30] developed a Gibbs sampler which samples topic assignments and table counts; later, Chen, Du, and Buntine [34] proposed a more effective algorithm that jointly samples topic assignments and ‘table indicators’ for each term. Table indicators are constraint variables that reconstruct table counts through summation. We use this block Gibbs sampler algorithm in our application of STM. The flavour of the Gibbs sampler is similar to the one of the standard LDA model, where topic distributions and document topic weights are ‘collapsed’ as a product of multinomials, allowing words to be sampled into topics directly. In the case of STM, this computation becomes more complex, because re-assigning words into topics has a potential knock-on effect on table indicators and table counts. Using numerical approximations, the conditional distributions can be computed, allowing Gibbs sampling of topic assignments and table indicators. See Appendix D for more inference details.

As in the case of LDA, the block Gibbs sampler algorithm does not explicitly sample topics ϕ , store-specific mixtures of objectives θ or transaction-specific mixtures of objectives ν . Instead, hidden variables are approximated using a posterior sample s of topic assignments and table counts. Then, hidden variables are approximated at each iteration by their conditional posterior means:

$$\hat{\theta}_{d,k}^s = E(\theta_{d,k}^s | \mathbf{t}^s, \boldsymbol{\alpha}) = \frac{\alpha_k + \sum_p t_{p,d,k}^s}{\alpha + \sum_{p,k} t_{p,d,k}^s} \tag{3}$$

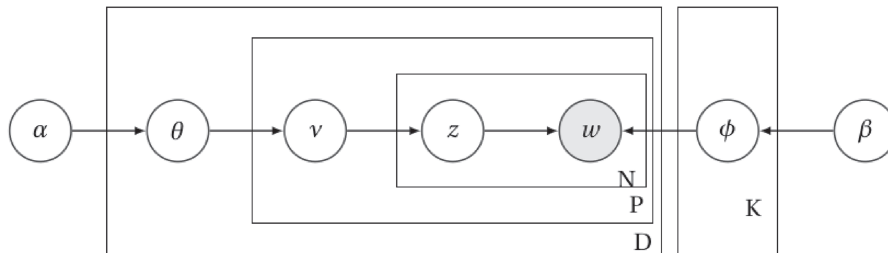


FIGURE 1 | STM graphical model in plate notation. Nodes denote random variables and edges denote dependencies. Unshaded node denote hidden random variables and shaded nodes denote observed random variables. Plates denote replication. The hidden variables are z topic assignments, θ store-specific topical mixtures, ν transaction-specific mixtures of objectives, ϕ topic distributions, α and β Dirichlet hyperparameters. Here K is the number of topics, D number of stores, P number of transactions, and N number of products.

$$\hat{p}_{p,d,k}^s = E(y_{p,d,k}^s | \mathbf{z}^s, \mathbf{t}^s, a, b) = \frac{N_{p,d,k}^s - a \times t_{p,d,k}^s}{b + N_{p,d}^s} + \theta_{d,k} \frac{\sum_k t_{p,d,k}^s \times a + b}{b + N_{p,d}^s} \quad (4)$$

$$\hat{\phi}_{k,v}^s = E(\phi_{k,v}^s | \mathbf{z}^s, \boldsymbol{\beta}) = \frac{\beta_v + N_{k,v}^s}{\beta + N_k^s} \quad (5)$$

where $\alpha = \sum_k \alpha_k$ and where $\beta = \sum_v \beta_v$.

3.2 | Summarising Topic Distributions

Summarising the posterior distribution of a topic model is challenging because the posterior distribution is often highly multi-modal, resulting in posterior samples that capture different shopping needs (in the text analysis context, these are called semantic modes). Although dominant topics, capturing common shopping needs, may consistently appear in every single MCMC iteration, topics with lower overall prevalence may only recognisably appear in some iterations [22, 23]. Thus, component-wise posterior averaging may merge topic distributions that respond to different semantic concepts.

In response, we follow the methodology described in [2] to construct a summary of topical modes using multiple posterior samples from various MCMC chains, similar in spirit to [35] and following clustering principles by Hennig et al. [36]. The rationale of this approach is that, in order to provide meaningful posterior summaries of topics, one needs to define a meaningful distance metric that can be used to assess whether two topics correspond to the same theme/concept, in this case a shopping need. To this end, the authors Vega Carrasco et al. [2] use the cosine similarity to compare similarity between topics, defined as

$$d(\phi_i, \phi_j) = \frac{\phi_i \cdot \phi_j}{\|\phi_i\| \|\phi_j\|} = \frac{\sum_{v=1}^V \phi_{iv} \phi_{jv}}{\sqrt{\sum_{v=1}^V \phi_{iv}^2} \sqrt{\sum_{v=1}^V \phi_{jv}^2}}$$

because it better captures relative ratios of weights of different products within each topic. The method [2] uses clustering to create groups of topics across different posterior samples that correspond to the same shopping need; these clusters can then be used to compute cluster-specific posterior summaries.

Specifically, the methodology inputs all topics across MCMC iterations and clusters them using a bottom-up hierarchical clustering approach. At each step, the algorithm finds the pair of clusters with the lowest cosine distance and merges clusters only if their topic distributions come from different samples. The algorithm keeps merging clusters up to a cosine distance threshold. Each resulting cluster is represented by the average topic distribution, named clustered topic, and by the number of topics gathered in the same cluster, named cluster size. Cluster size is a measure of recurrence and denotes (un)certainty, that is, a topic that has occurred in every posterior sample is highly recurrent, showing no uncertainty. Although use this approach was used within a standard LDA model [2], the methodology is naturally applicable to STMs as well.

3.3 | Evaluation of Clustered Topics

Depending on the cosine distance threshold, the clustering algorithm produces a set of clustered topics, which can be

selected according to their recurrence. Thus, multiple subsets of clustered topics can be formed by varying cosine distance threshold and recurrence (setting a minimum cluster size). We evaluate each subset of clustered topics on four aspects: generalisation or predictive power of a subset of topics, coherence of individual topics, the distinctiveness of a topic with respect to the other topics in the same posterior sample, and credibility of a topic with respect to the topics from other posterior samples.

Topic coherence, distinctiveness and credibility are measured as described in [2]. Model generalisation, however, is measured by the perplexity of unseen transactions given topics, store-specific mixtures of objectives and PDP parameters:

$$\text{Perplexity} = - \frac{\log P(\mathbf{w}'_d | \Phi, \theta_d, a, b)}{N'} \quad (6)$$

where \mathbf{w}'_d is a set of products in a held-out transaction at store d , N' is the number of products in \mathbf{w}'_d , $\Phi = [\phi_1, \phi_2, \dots, \phi_K]$ the set of inferred topics, θ_d is the store-specific mixtures of objectives associated to store d , a and b are the PDP parameters. We aim to select a subset of clustered topics that shows low perplexity, gathering topics that are coherent, distinctive and credible (low uncertainty).

The quality of the clustered topics is influenced by both the number of topics specified in the sampler and the choice of hyperparameters. In this article, we chose the hyperparameters and the number of topics through an exploration of various combinations. For a detailed discussion on model evaluation and model selection, we refer readers to our earlier paper [2].

4 | Gaussian Process Regression

According to Tobler's first law of geography [37]: 'everything is related to everything else, but near things are more related than distant things'. Thus, we expect that nearby stores show similar shopping patterns and that some specific patterns may be limited to particular geographical areas. STM does not take into account store location or proximity between stores. Although a topic model that simultaneously accommodate store hierarchy over transactions and store location would be mathematically possible, it would be computationally prohibitive at the level of resolution of interest. Instead, we use the summarised posterior distributions of topics obtained from STM and take a spatial modelling approach to capture their geographical structure and regional behaviour.

4.1 | Model

A linear regression with spatial Gaussian process errors is defined as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\eta} + \boldsymbol{\varepsilon} \quad (7)$$

where \mathbf{Y} is the dependent variable, \mathbf{X} is the matrix of p covariates associated with locations $\mathbf{s}_1, \dots, \mathbf{s}_n$, $\boldsymbol{\beta}$ is a p -dimensional fixed effect, $\boldsymbol{\eta}$ is a spatial process, which captures spatial residual, and $\boldsymbol{\varepsilon}$ is an independent process, which models pure error, also known as the *nutget* effect.

The spatial process $\eta(\mathbf{s}_1), \dots, \eta(\mathbf{s}_n)$ is distributed as a zero-mean Gaussian process $GP(0, C_\eta)$ with positive definite covariance matrix C_η . Residuals $\varepsilon(\mathbf{s}_1), \dots, \varepsilon(\mathbf{s}_n)$ are assumed *iid* with $\varepsilon(\mathbf{s}_i) \sim N(0, \sigma^2)$. Thus, observations are distributed as:

$$\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \Sigma) \quad (8)$$

where $\Sigma = C_\eta + \sigma^2 I$.

Here, we use the positive definite square exponential covariance function,

$$C_\eta(\mathbf{s}_i, \mathbf{s}_j | \alpha, \rho) = \alpha^2 \exp\left(-\frac{\text{dist}(\mathbf{s}_i, \mathbf{s}_j)^2}{2\rho^2}\right) \quad (9)$$

where parameters α and ρ control the amplitude and length-scale of the spatial dependence, respectively. $\text{dist}(\mathbf{s}_i, \mathbf{s}_j)$ is a measure of distance between locations.

4.2 | Methods

The GPR specified in Equation (7) is fitted using Stan [38]. Stan facilitates Bayesian inference by gradient-based sampling techniques such as Hamiltonian Monte Carlo methods [39] and variational inference [40]. In our study, the inference is computed by the default Stan algorithm No-U-Turn Sampler (NUTS) [41]. NUTS is an extension of the Hamiltonian Monte Carlo (HMC) algorithm that effectively explores the parameter space by avoiding retaking previously sampling paths in a U-turn style.

4.3 | Predictions

Predicted topic probabilities $\mathbf{Y}^* = [Y^*(\mathbf{s}_1), \dots, Y^*(\mathbf{s}_n)]$ at new locations $\mathbf{s}_1^*, \dots, \mathbf{s}_n^*$ are distributed as:

$$\begin{aligned} & \mathbf{Y}^* | \mathbf{Y}, \boldsymbol{\beta}, \Theta, \mathbf{X}^*, \mathbf{X} \\ & \sim N(\mathbf{X}^* \boldsymbol{\beta} + \Sigma_{21} \Sigma_{11}^{-1} (\mathbf{Y} - \mathbf{X} \boldsymbol{\beta}), \Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12}) \end{aligned} \quad (10)$$

where \mathbf{X}^* is the matrix of p covariates at the new locations. Here, Σ_{11} is the covariance matrix of $\mathbf{s}_1, \dots, \mathbf{s}_n$ locations, $\Sigma_{12} = \Sigma_{21}$ the covariance matrix between $\mathbf{s}_1, \dots, \mathbf{s}_n$ and $\mathbf{s}_1^*, \dots, \mathbf{s}_n^*$, and Σ_{22} , covariance matrix of $\mathbf{s}_1^*, \dots, \mathbf{s}_n^*$. The first quantity in the expected topic probabilities $E(\mathbf{Y}^*)$ are computed is obtained by multiplying the design matrix by the fixed effects as in multiple linear regression. The second quantity pulls the expected value at a new store towards the values of the nearby stores if spatial dependence is significant.

5 | Identifying Regional Grocery Topics

We analyse grocery transactions from a major retailer in the UK. Transactions are sampled randomly, covering 100 nationwide superstores between September 2017 and August 2018. Transactions with fewer than 3 products are filtered out because they inflate the dataset without providing additional meaningful information. The training data set contains 36,000 transactions and a total of 392,840 products and the test data set contains

3600 transactions and a total of 38,621 products. Transactions contain 10 products on average. The product assortment contains 10,000 products, which are the most monthly frequent, ensuring the selection of seasonal and non-seasonal products. We count unique products in transactions, disregarding the quantities of repetitive products. For instance, 5 loose bananas count as 1 product (loose banana). We do not use an equivalent of stop words list (highly frequent terms), as we consider that every product or combination of them tell different customer needs. We disregard transactions with fewer than 3 products assuming that smaller transactions do not have enough products to exhibit a regional topic. No personal customer data were used for this research.

5.1 | STM Posterior Summary

We explore STM with 100 topics to capture as many topics as possible without making inference too computationally prohibited. As shown in [2], a topic model with 100 topics identifies a variety of customer behaviours in the domain of our application. Exploring STM with a smaller or larger number of topic is out the scope of this article.

We use symmetric priors with hyperparameters $\alpha_k = 1000/K$ and $\beta_v = 0.01$, and PDP hyperparameters $b = 3.0$ and $a = 0.5$. We run four MCMC chains for 100,000 iterations with a burn-in of 80,000 iterations, samples were recorded every 5000 iterations, obtaining 20 thinned posterior samples (five samples for each chain). MCMC trace plots are presented in Appendix A, where the convergence is satisfactory.

Posterior topic distributions are summarised by clustering a bag of 2000 topics obtained from the aforementioned 20 posterior samples. As shown in Appendix C, we observe that the subset formed with a minimum cluster size 10 (which represent 50% of the samples) and a cosine distance threshold ≥ 0.35 show greater coherence, credibility and generalization, concurring with [2]. Based on these results, we choose this subset which contains 104 clustered topics.

To obtain store-specific mixtures of shopping objectives across the identified 104 clustered topics, the STM is re-fitted using these 104 clustered topic distributions, which are held fixed during the re-fitting process. The sampler follows the steps described in Section 3.1.1, but only Equations (3) and (4) are updated. A MCMC chain runs with a burn-in period of 1000 iterations, recording posterior samples with a thin of 500 iterations. The MCMC trace plot in Appendix B shows satisfactory convergence. We collect 30 posterior samples which are then averaged to estimate store-specific mixtures of objectives for 500 stores across the UK.

5.2 | Interpreting Topic Distributions

We interpret six out of the 104 clustered topics as they capture a clear, interpretable regional pattern. A few more topics, along with their geographical distribution, are shown in Appendix F. The majority of the remaining topics either show a ubiquitous distribution across the UK or do not have a clear, unique interpretation in terms of the underlying shopping objectives. We interpret topics by analysing the product descriptions of the

15 products with the largest probabilities. Topics are manually named after the regional pattern or customer preference reflected on the product descriptions. Note that the six illustrated topics appeared consistently across the 20 posterior samples (size = 20), indicating low posterior uncertainty.

Product descriptions in Figure 2a–c suggest foods supplied locally and local brands associated to Scotland, Northern Ireland and Wales. For instance, the Scottish topic includes ‘Scottish-branded skinless sausages’ and ‘Scottish-branded potato scones’, the Northern Irish topic shows the ‘North Ireland semi-skimmed milk’, ‘white potatoes packed in North Ireland’, and the Welsh topic contains ‘Welsh jacket potatoes’ and ‘Welsh-branded bread’. Hence, we name the Scottish topic, Northern Irish topic and Welsh topic after the nationality that their product descriptions suggest.

Figure 2d,e show a variety of products such as types of milk, types of bread, fruits and vegetables and so forth. Close inspection of product descriptions such as ‘oven bottom muffin’, ‘fruit teacake’, and ‘potato and meat pie’ in Figure 2d, and ‘pork pies’ and ‘scotch eggs’ in Figure 2e may reveal a regional topic when regional expertise is available. Since these product descriptions do not provide interpretations that can be directly associated with specific regions, we momentarily name these topics ‘Mixed basket I’ and ‘Mixed basket II’. Figure 2f shows ‘organic’ quality foods, indicating a specific customer preference, however, the topic does not suggest any specific regional pattern.

Interpreting topic descriptions is not sufficient to identify geographically driven shopping motivations, reinforcing the need for exploring store-specific mixtures of objectives.

5.3 | Mapping Mixtures of Objectives

Interpreting product descriptions may reveal the existence of regional topics, that is, the Northern Irish/ Scottish/ Welsh topic. Topic interpretations may dismiss regional topics that exhibit products that are not directly linked with specific areas. Thus, we visualise the store-specific topic probabilities at the store’s location, aiming to find topics with a regional pattern. We link store postcodes with location coordinates through querying stores’ postcodes in the lookup table from the Office for National Statistics [42]. Figure 3 shows the topic probabilities of the six clustered topics mapped across the UK.

Figure 3a–c clearly confirm that the Scottish, Northern Irish and Welsh topics are more likely in their respective constituent countries. More interestingly, Figure 3c shows the prevalence of the Welsh topic over neighbouring regions. Figure 3d shows high topic probabilities concentrated in the North West and surrounding regions, and Figure 3e shows high topic probabilities in the central and southern English regions. We rename both topics as North and Centre and South and Midlands due to their cross-regional prevalence. Figure 3f, which maps the Organic topic, shows significant probabilities concentrated in London.

In comparison to the Scottish, Northern Irish and Welsh topics, interpretations of the most probable topics in the North and Centre, South and Midlands and Organic topics do not easily suggest a geographical pattern. Mapping the store-specific topic probabilities aids the analysis and identification of topics with spatial patterns.

(a) Scottish	(b) Northern Irish	(c) Welsh
NPMI = 0.26 Size = 20	NPMI = 0.31 Size = 20	NPMI = 0.18 Size = 20
0.0904 BRITISH S/SKIMMED MILK 2.272L, 4 PINTS	0.0851 NORTHERN IRELAND S/SKIMMED MILK 2 LTR	0.0873 BRITISH S/SKIMMED MILK 2.272L, 4 PINTS
0.0449 BRITISH WHOLE MILK 2.272L, 4 PINTS	0.0327 NORTHERN IRELAND WHOLE MILK 2 LTR	0.0284 BRITISH WHOLE MILK 2.272L, 4 PINTS
0.0331 XXX TOASTIE SLICED WHITE BREAD 800G	0.0324 BANANAS LOOSE	0.0233 RIPE BANANAS 5 PACK
0.0278 SC-XXX CRISPY MORNING ROLL	0.0263 NORTHERN IRELAND S/SKIMMED MILK 3 LTR	0.0226 XXX WELSH WHITE POTATO 2.5KG
0.0272 XXX MEDIUM SLICED WHITE BREAD 800G	0.0227 XXX SOFT WHITE MEDIUM BREAD 800G	0.0175 WHITE TOASTIE THICK BREAD 800G
0.0218 BRITISH S/SKIMMED MILK 1.13L, 2 PINTS	0.0192 NORTHERN IRELAND S/SKIMMED MILK 1 LTR	0.0172 XXX SPREAD 500 G
0.0183 RIPE BANANAS 5 PACK	0.0175 RIPE BANANAS 5 PACK	0.0162 WE-XXX WHITE THICK SLICED LOAF 800G
0.0163 XXX SLIGHTLY SALTED SPREADABLE 500G	0.0169 WHITE POTATOES 2KG PACKED NI	0.0155 CLOSED CUP MUSHROOMS 300G
0.0153 SC-XXX POTATO SCONES 6 PK	0.0159 MEDIUM FREE RANGE EGGS 6 PACK	0.0134 XXX LAGER 18X440ML
0.0148 SC-XXX SCOTTISH PLAIN WHT BRD 800G	0.0146 NI-XXX PANCAKES 6 PACK	0.0129 XXX WELSH BABY POTATO 1KG
0.0125 SMOKED BACK BACON RASHERS 300G	0.0146 NI-XXX NAVAN POTATOES 2KG	0.0125 WE-XXX JACKET POTATOES 700G
0.0118 XXX TOASTIE WHITE SMALL BREAD 400G	0.0129 BUNCHED SPRING ONIONS 100G	0.0114 FREE RANGE EGGS MEDIUM 6 PK
0.0116 WAFER THINHONEY ROAST HAM SLICES 125G	0.0126 PANCAKES 8PK	0.0109 WE-XXX WHITE MEDIUM SLICED LOAF 800G
0.0106 SC-XXX MACARONI CHEESE 250G (L)	0.0125 CLOSED CUP MUSHROOMS 300G	0.0108 SMOKED BACK BACON RASHERS 300G
0.0105 SC-XXX ORIGINAL SMOKED PORK SAUSAGE 200G	0.0124 BROWN ONIONS 3PK 385G	0.0101 XXX ORANGE JUICE SMOOTH 1.6 LTR
(d) Mixed basket I	(e) Mixed basket II	(f) Organic
NPMI = 0.26 Size = 20	NPMI = 0.28 Size = 20	NPMI = 0.35 Size = 20
0.0594 BRITISH S/SKIMMED MILK 2.272L, 4 PINTS	0.0726 BRITISH S/SKIMMED MILK 2.272L, 4 PINTS	0.0426 ORGANIC FAIRTRADE BANANAS 6 PACK
0.0403 XXX TOASTIE SLICED WHITE BREAD 800G	0.0524 BRITISH S/SKIMMED MILK 1.13L, 2 PINTS	0.0404 ORGANIC CARROTS 700G
0.0318 XXX CRUMPETS 6 PACK	0.0257 CLOSED CUP MUSHROOMS 300G	0.0262 ORGANIC BRITISH S/SKIMMED MILK 4 PINTS
0.0171 XXX MEDIUM SLCD WHT BRD 800G	0.0252 WHITE BAGUETTE 400G	0.0243 MIXED SIZED ORGANIC EGGS 6 PACK
0.0169 NE-XXX OVEN BOTTOM MUFFINS 6 PACK	0.0223 BANANAS LOOSE	0.0183 ORGANIC GALA APPLES 630G
0.0167 BRITISH WHOLE MILK 2.272L, 4 PINTS	0.018 XXX CRUMPETS 6 PACK	0.0164 ORGANIC BRITISH S/SKIMMED MILK 2 PINT
0.0163 IRISH-XXX 8 THICK PORK SAUSAGES 454G	0.0179 XXX ORIGINAL SPREAD 500 G	0.0146 ORGANIC BROCCOLI 300G
0.0154 XXX SLIGHTLY SALTED SPREADABLE 500G	0.0144 6 HOT CROSS BUNS	0.0136 ORGANIC WHITE POTATOES 1.5KG
0.0147 PREMIUM JACKET POTATOES 4 PACK	0.0136 TIGER BAGUETTE 400G	0.0132 ORGANIC UNSALTED BTRR 250G
0.014 UNSMOKED THICK CUT BACK BACON 300G	0.0134 XXX SOFT WHITE THICK BREAD 800G	0.0125 RIPE & READY TWIN PACK AVOCADOS
0.0136 UNSMOKED BACK BACON RASHERS 300G	0.0122 XXX SALTED SPREADABLE 500G	0.0124 ORGANIC HOUMOUS 200G
0.0135 WHITE BATON	0.0114 XXX MATRURE CHEDDAR CHEESE 550 G	0.0121 READY TO EAT LARGE AVOCADOS EACH
0.0131 XXX TOASTIE WHITE SMALL BREAD 400G	0.0114 BRITISH SALT BLOCK BUTTER 250G	0.0111 ORGANIC SMALL BANANAS 6 PACK
0.0126 XXX WHITE SLICED SANDWICH ROLLS 6 PACK	0.011 PREMIUM 12 PORK BRITISH CHIPOLATAS 375G	0.0108 ORGANIC BRITISH WHOLE MILK 4 PINTS
0.0108 EGG CUSTARD TARTS 4 PACK	0.0103 BRITISH CRUMBED HAM SLICES 125 G	0.0106 RASPBERRIES 150G

FIGURE 2 | Most probable products in grocery regional topics. Each topic is interpreted using the 15 products with the largest probabilities. Probabilities and products are sorted in descending order. General brand names have been replaced by XXX. Local brands in North Ireland, Scotland, Wales and North of England have been replaced by NI-XXX, SC-XXX, WE-XXX, NE-XXX. NPMI and size are measures of topic coherence and recurrence.

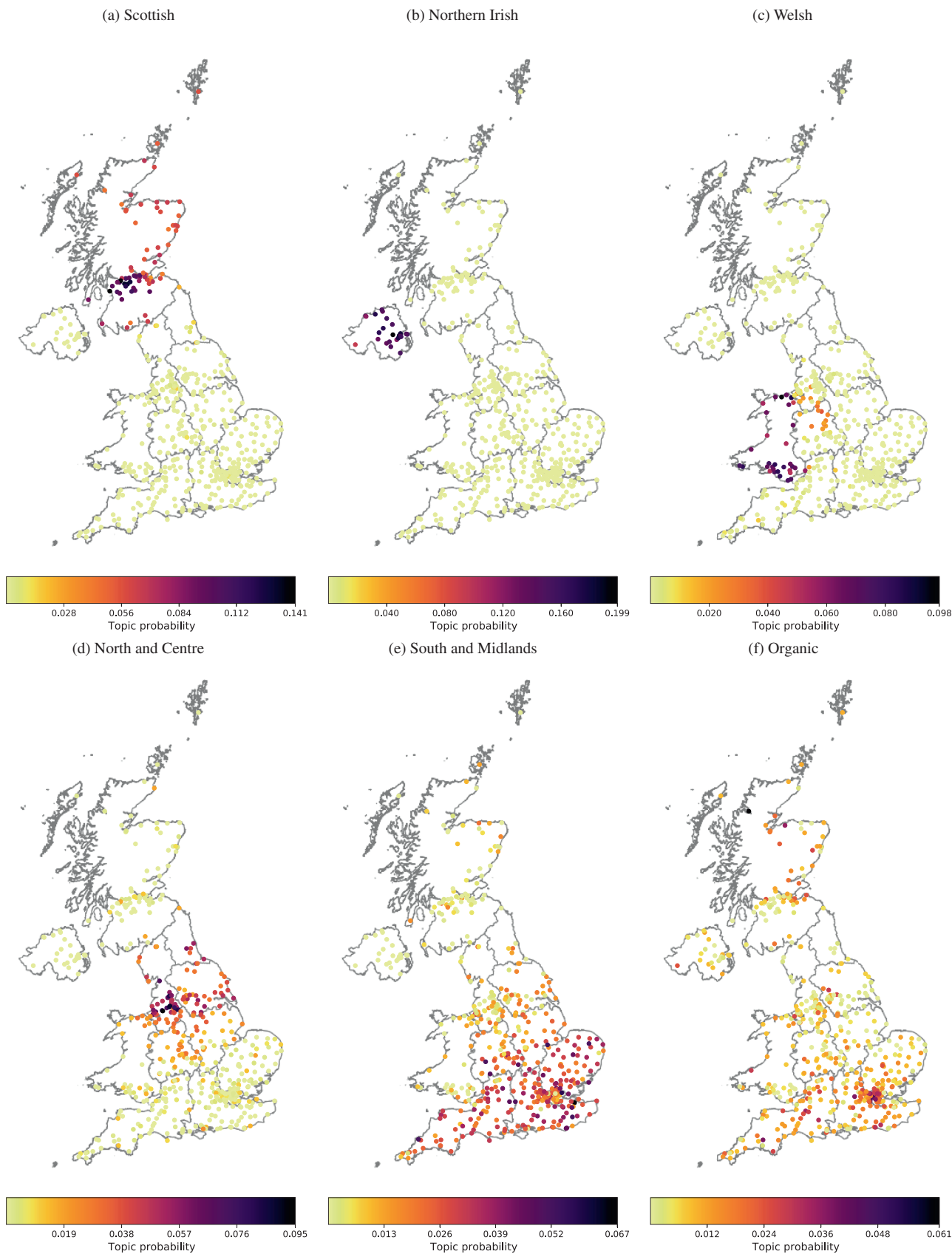


FIGURE 3 | Topic probabilities of clustered grocery topics in the UK. Purple and yellow points reflect the largest and smallest topic probabilities, respectively.

5.4 | STM Versus LDA

STM shows two advantages over LDA. First, STM provides topical summaries for stores, by including the store hierarchy above transactions. Second, and less obvious, STM discovers topics that are relevant within their store context. In comparison, LDA finds products that are frequently bought together across all transactions. Thus, a product combination that is only frequent in few stores may not be shown among LDA topics. The ability to capture store-specific topics is key to our subsequent spatial modelling analysis.

We compare the 104 STM clustered topics (HC-STM-100) against the posterior summaries of the LDA model with 100 and 200 topics. The posterior summaries of LDA were obtained using the same training data and following the clustering methodology in [2]. The posterior summary of LDA with 100 topics (HC-LDA-100) gathers 96 clustered topics and the posterior summary of LDA with 200 topics (HC-LDA-200) gathers 198 clustered topics.

Figure 4a shows the cosine similarity between (HC-STM-100) 104 clustered topics and (HC-LDA-100) 96 clustered topics. Clustered topics are ordered to visualise high similarities in the diagonal. As observed, the majority of clustered topics are identified in both models, STM and LDA, with high cosine similarity >0.7 . Figure 5a shows that 70% of the (HC-STM-100) clustered topics are found among HC-LDA-100 clustered topics, and 85% of the HC-LDA-100 clustered topics are found among the HC-STM-100 clustered topics with high similarity. For instance, the Northern Irish topic is found in both models with high cosine similarity (0.97). As depicted in Figure 6a, Northern Ireland related products rank in the top 15 products in both topics. The Organic topic

was also found among HC-LDA-100 clustered topics with high cosine similarity (0.95).

We also compare the 104 STM clustered topics against the 198 LDA clustered topics obtained from summarising LDA posterior samples of 200 topics. This comparison allows identifying the regional topics that were not caught in LDA samples with 100 topics. For instance, the Scottish topic described in Figure 2a, is not found in the HC-LDA-100 subset, but it is found in the HC-LDA-200 subset with a cosine similarity of 0.83. As observed in Figure 4b, the majority of the 104 clustered topics are found among the (HC-LDA-200) 198 LDA clustered topics with high cosine similarity (>0.7). However, Figure 5b shows that there are still some STM clustered topics that do not match with any of the LDA clustered topic with high similarity. For example, the Welsh topic described in Figure 2c, is not found in either of the two subsets of LDA clustered topics. The Welsh topic and the closest clustered topic in HC-LDA-200 (with 0.67 cosine similarity) are listed in Figure 6b; as observed, few products are shared by the topics but Welsh products are not described in both topics. The North and Centre topic and the South and Midlands topic were not found among the HC-LDA-200 clustered topics either. Perhaps, these regional topics would appear among posterior samples of a larger LDA model, that is, LDA with 300 topics; however, increasing the model complexity is not only more computationally expensive but also less efficient since topics tend to show less distinctiveness [2].

In summary, three out of the six regional topics are identified by STM and the other three regional topics are identified by both STM and LDA models. One of these topics was captured by a larger LDA model. STM shows its strength over LDA by identifying more regional topics.

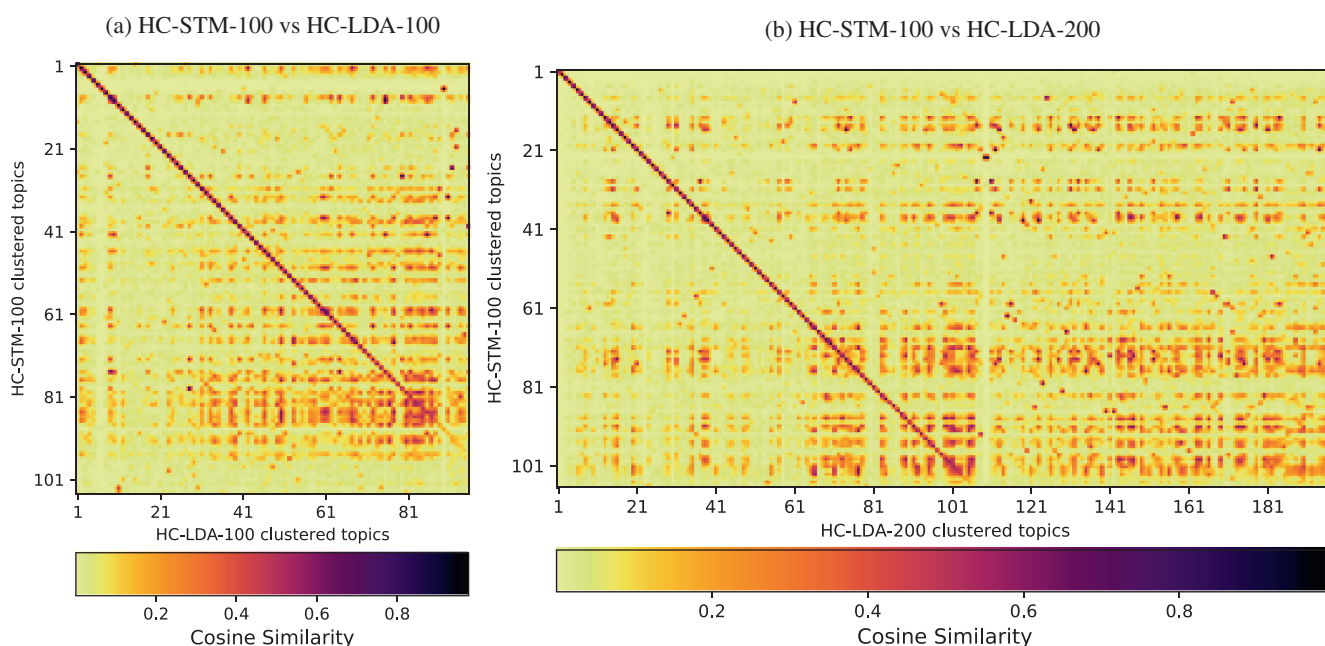
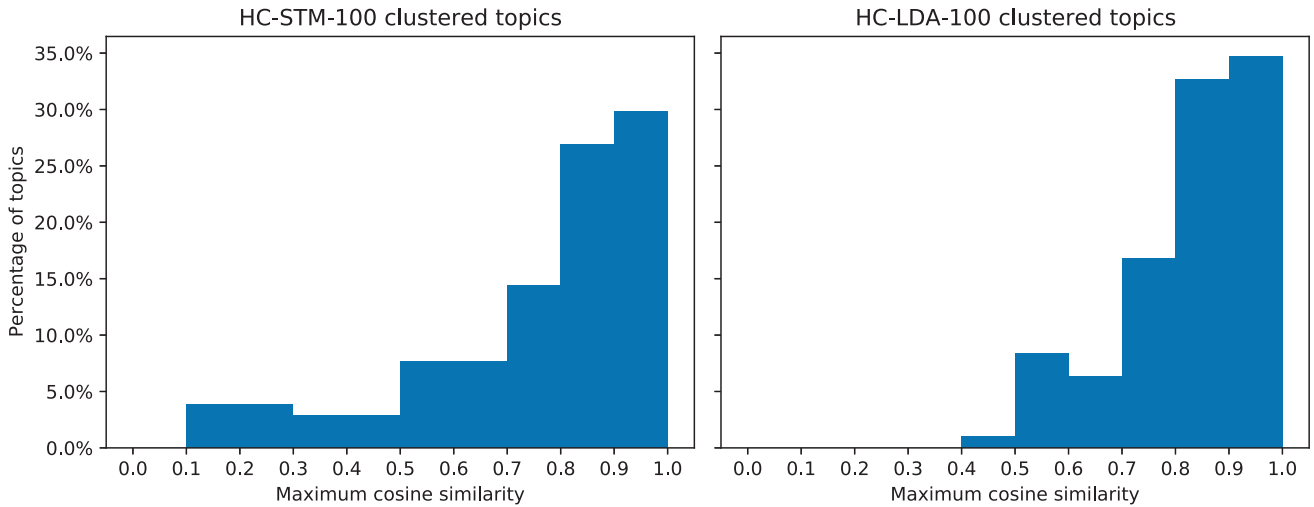


FIGURE 4 | Cosine similarity between clustered topics obtained from posterior summaries of STM with 100 topics and LDA with 100 and 200 topics. Topics have been aligned following a greedy algorithm that at each step searches and pairs topics (that have not been paired) with the highest cosine similarity.

(a) HC-STM-100 vs HC-LDA-100



(b) HC-STM-100 vs HC-LDA-200

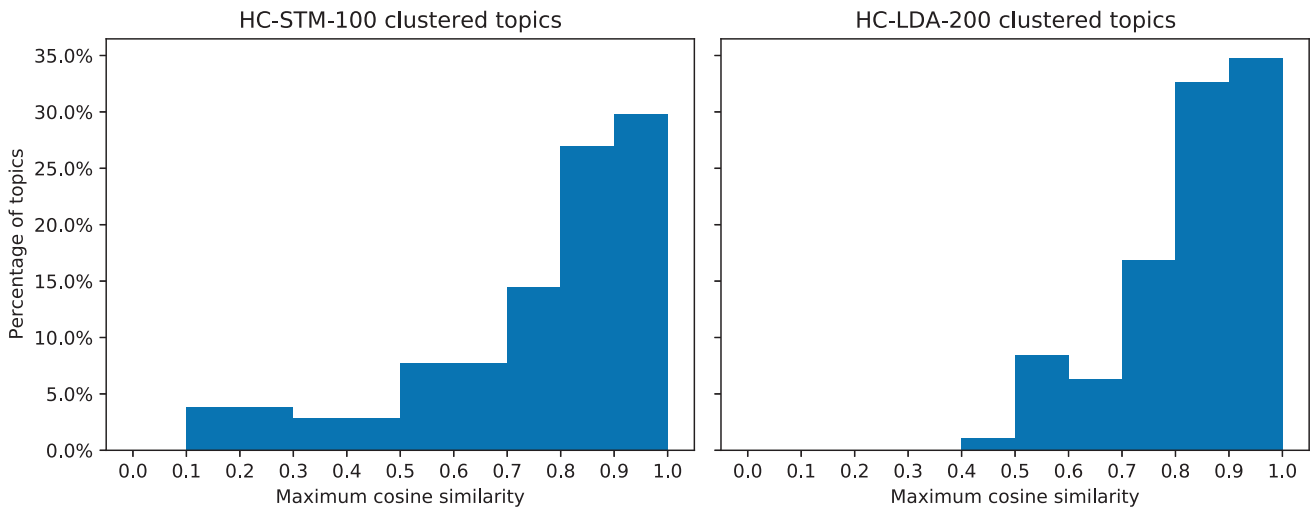


FIGURE 5 | Distributions of the maximum cosine distance obtained from each cosine similarity matrix in this figure. (a) Plots maximum cosine distances between clustered STM topics (HC-STM-100) against the posterior summary of LDA with 100 topics (HC-LDA-100) (left); and from HC-LDA-100 to HC-STM-100 (right). (b) Plots maximum cosine distances between HC-STM-100 against the posterior summary of LDA with 200 topics (HC-LDA-200) (left); and from HC-LDA-200 to HC-STM-100 (right).

6 | Modelling Regional Prevalence

We aim to model topic probabilities across stores in the UK by constructing a linear model with fixed effects associated with the constituent countries of the UK (Wales, Scotland, Northern Ireland) and the nine English regions, and imposing spatial dependency through a Gaussian process that captures residual spatial association as defined in Equation (7). In this manner, we can quantify the significance of a topic to a region or constituent country. This administrative division was chosen assuming that each country and region would broadly show differences in customer behaviour. Analysis over other subdivisions is possible, but it is out of the scope of this article.

The dependent variable \mathbf{Y}_k is the logit transformation of the store-specific k th topic probabilities $[\hat{\theta}_{s_1,k}, \hat{\theta}_{s_2,k}, \dots, \hat{\theta}_{s_n,k}]$,

given by:

$$\mathbf{Y}_k = \text{logit}([\hat{\theta}_{s_1,k}, \hat{\theta}_{s_2,k}, \dots, \hat{\theta}_{s_n,k}]) \quad (11)$$

where each $\hat{\theta}_{s_i,k}$ is the average probability over 30 posterior samples of the k th topic at store location s_i from Section 5.1. For simplicity, we assume independence among topic probabilities and model each topic separately, that is, for each topic, a linear model is constructed. However, topic probabilities of a topical mixture are not independent of each other since they need to sum to 1.

The logit transformation not only avoids predicting nonsensical values (i.e., topic probabilities >1 or <0), but also aids the visualisation of topic probabilities that cannot be appreciated in the original scale. For instance, Figure 7 (left panel) highlights stores in the South West that are not noticed in Figure 3c.

(a) The Northern Irish topic in STM and LDA

Clustered STM	Clustered LDA
. NORTHERN IRELAND S/SKIMMED MILK 2 LTR	. NORTHERN IRELAND S/SKIMMED MILK 2 LTR
. NORTHERN IRELAND WHOLE MILK 2 LTR	. BANANAS LOOSE
. BANANAS LOOSE	. NORTHERN IRELAND WHOLE MILK 2 LTR
. NORTHERN IRELAND S/SKIMMED MILK 3 LTR	. NORTHERN IRELAND S/SKIMMED MILK 3 LTR
. XXX SOFT WHITE MEDIUM BREAD 800G	. XXX SOFT WHITE MEDIUM BREAD 800G
. NORTHERN IRELAND S/SKIMMED MILK 1 LTR	. RIPE BANANAS 5 PACK
. RIPE BANANAS 5 PACK	. WHITE POTATOES 2KG PACKED NORTHERN IRELAND
. WHITE POTATOES 2KG PACKED NORTHERN IRELAND	. NI-XXX COUNTRYNAVAN POTATOES 2KG
. MEDIUM FREE RANGE EGGS 6 PACK	. NI-XXX PANCAKES 6 PACK
. NI-XXX PANCAKES 6 PACK	. CLOSED CUP MUSHROOMS 300G
. NI-XXX COUNTRYNAVAN POTATOES 2KG	. PANCAKES 8PK
. BUNCHED SPRING ONIONS 100G	. SALAD TOMATOES 6 PACK
. PANCAKES 8PK	. NORTHERN IRELAND S/SKIMMED MILK 1 LTR
. CLOSED CUP MUSHROOMS 300G	. MEDIUM FREE RANGE EGGS 6 PACK
. BROWN ONIONS 3PK 385G	. JAFFA CLEMENTINE OR SWEETEASY PEELER 600G

(b) The Welsh topic in STM and its most similar topic in LDA

Clustered STM	Clustered LDA
. BRITISH S/SKIMMED MILK 2.272L, 4 PINTS	. BRITISH S/SKIMMED MILK 2.272L, 4 PINTS
. BRITISH WHOLE MILK 2.272L, 4 PINTS	. BRITISH WHOLE MILK 2.272L, 4 PINTS
. RIPE BANANAS 5 PACK	. UNSMOKED BACK BACON RASHERS 300G
. XXX WELSH WHITE POTATO 2.5KG	. CLOSED CUP MUSHROOMS 300G
. WHITE TOASTIE THICK BREAD 800G	. SMOKED BACK BACON RASHERS 300G
. XXX SPREAD 500 G	. XXX 8 THICK PORK SAUSAGES 454G
. WE-XXX WHITE THICK SLICED LOAF 800G	. XXX TOASTIE SLICED WHITE BREAD 800G
. CLOSED CUP MUSHROOMS 300G	. UNSMOKED THICK CUT BACK BACON 300G
. XXX LAGER 18X440ML	. SMOKED THICK CUT BACK BACON 300G
. XXX WELSH BABY POTATO 1KG	. XXX MEDIUM SLICED WHITE BREAD 800G
. WELSH JACKET POTATOES 700G	. MARIS PIPER POTATOES 2.5KG
. FREE RANGE EGGS MEDIUM 6 PK	. XXX MIXED SIZED EGGS 10 PACK
. XXX WHITE MEDIUM SLICED LOAF 800G	. XXX PUDDING 4 SLICES 230G
. SMOKED BACK BACON RASHERS 300G	. BRITISH S/SKIMMED MILK 1.13L, 2 PINTS
. XXX ORANGE JUICE SMOOTH 1.6 LTR	. WHITE TOASTIE THICK BREAD 800G

FIGURE 6 | Comparison of topics identified in STM and LDA posterior samples. Highlighted products appear in both topics. While the Northern Irish topic is clearly identified by both models, the Welsh topic is only found by the STM model.

The covariate matrix \mathbf{X} is defined by dummy variables responding to the constituent countries: ‘North Ireland’, ‘Scotland’, ‘Wales’; and the English regions: ‘North East’, ‘North West’, ‘Yorkshire and the Humber’, ‘East Midlands’, ‘West Midlands’, ‘South West’, ‘South East’, and ‘East Anglia’, where ‘London’ is the reference category.

The spatial distance $\text{dist}(\mathbf{s}_i, \mathbf{s}_j)$, which define covariance between stores $C_\eta(\mathbf{s}_i, \mathbf{s}_j)$, is calculated by first finding the latitude-longitude coordinates associated with the store’s postcode, second computing the distance between pair of coordinates using the Haversine formula [43]. The Haversine formula provides accurate approximations of distance for locations over large areas. Postcode coordinates are queried from the postcode lookup table from the Office for National Statistics [42]. Spatial distance is measured in kilometres.

We complement the Bayesian hierarchical model with weakly informative priors: $\sigma^2 \sim \text{halfN}(0, 1)$, $\beta \sim N(0, 10)$; $\alpha \sim N(0, 2)$, and $\rho \sim IG(2, 50)$.

Parameters of the GPR are estimated with Stan, using 2 MCMC chains which run for 2000 iterations, 1000 burn-in iterations, and a thin of five iterations. Convergence of MCMC chains is satisfactory with scale factor reduction $\hat{R} = 0.998$.

6.1 | Prevalence of Regional Behaviours in the UK

We take a closer look at regional behaviours, drawing parallels to the 2012 Food Survey study published by the Department for Food and Rural Affairs [44]. Table 1 shows posterior summaries of the GPR. The intercept can be interpreted as how likely (in logit scale) a topic is at a store in London and vice versa. Positive average coefficients indicate that the topic is more likely than in London. Average coefficients that are highlighted in red correspond to non-zero 95% credible intervals with $0 >$ upper bound, and bold average coefficients correspond to non-zero 95% credible intervals with $0 <$ lower bound.

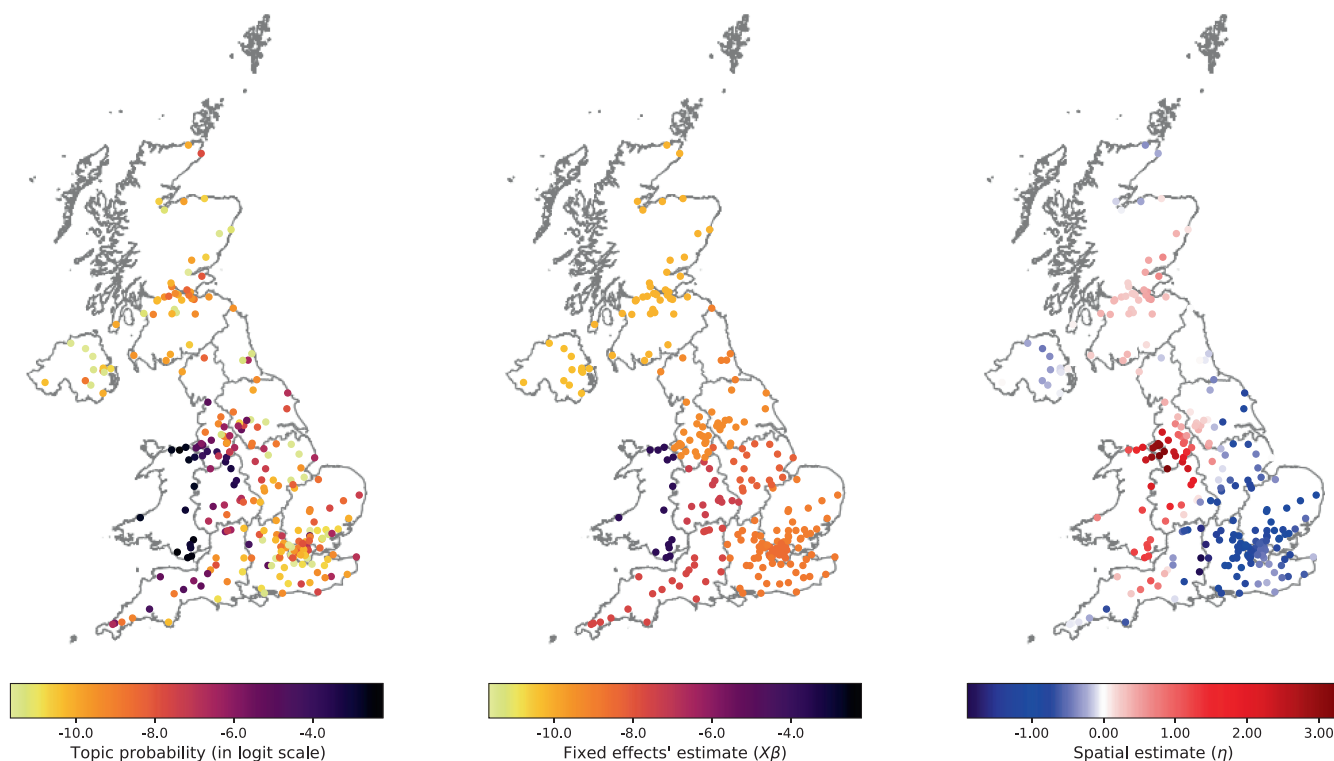


FIGURE 7 | Welsh topic: (left panel) observed topic probabilities in logit scale; (central panel) probability estimates (in logit scale) using only fixed effects; (right panel) spatial residuals captured by the Gaussian process.

TABLE 1 | Regression parameters for regional topics **Red/bold** mean estimates for coefficients with non-zero credibility intervals that decrease/increase the topic probability, respectively.

Parameter	Northern Irish		Scottish		Welsh		North and Centre		South and Midlands		Organic	
	Avg.	SE	Avg.	SE	Avg.	SE	Avg.	SE	Avg.	SE	Avg.	SE
Intercept	-10.4	0.02	-9.52	0.03	-8.9	0.04	-6.34	0.04	-4.42	0.02	-4.62	0.05
Northern Ireland	8.67	0.03	-0.72	0.04	-1.44	0.07	-4.11	0.05	-5.77	0.03	-1.25	0.06
Scotland	0.19	0.02	6.84	0.04	-1.12	0.05	-1.93	0.04	-1.82	0.03	-1.34	0.06
Wales	-0.4	0.03	-0.57	0.03	5.63	0.07	0.39	0.04	-2.27	0.03	-1.27	0.06
North West	0.15	0.03	1.54	0.04	0.1	0.08	3.3	0.05	-0.99	0.03	-1.91	0.06
North East	-0.86	0.04	3.27	0.05	-0.33	0.08	3.05	0.06	-1.25	0.04	-2.5	0.07
Yorkshire	0.04	0.03	1.08	0.04	-0.43	0.05	2.98	0.06	-0.43	0.03	-1.68	0.05
West Midlands	-0.15	0.02	-0.24	0.03	1.89	0.07	1.95	0.05	0.26	0.03	-1.01	0.05
East Midlands	-0.47	0.03	0.68	0.04	0.67	0.05	1.45	0.05	0.31	0.06	-1.47	0.05
East Anglia	-0.27	0.02	-0.28	0.03	-0.38	0.05	-0.31	0.04	0.99	0.02	-1.03	0.05
South East	-0.21	0.2	0.56	0.03	-0.25	0.04	-1.07	0.04	0.66	0.02	-0.51	0.05
South West	-0.26	0.2	-0.1	0.03	1.26	0.05	-0.64	0.04	0.71	0.03	-0.02	0.05
Length-scale ρ	63.85	5.95	92.07	19.95	55.31	1.32	51.32	15.53	50.23	3.84	34.67	3.13
Amplitude α	0.13	0.01	0.3	0.03	1.04	0.01	0.74	0.02	0.23	0.01	0.86	0.02
σ	0.78	0.01	1.38	0.01	1.43	0.01	1.37	0.01	1.15	0.01	1.58	0.01

Unsurprisingly, the Scottish, Northern Irish and Welsh topics show positive average coefficients with non-zero credibility intervals for the respective constituent countries. This indicates that their topic probability largely increases for stores in Scotland, North Ireland and Wales, respectively.

Interestingly, Wales's and Scotland's neighbouring regions show positive average coefficients with non-zero credibility intervals, that is, North East and North West to the Scottish topic and West Midlands and South West to the Welsh topic. As shown in Figure 7 (central panel), probability estimates (in logit scale)

of the Welsh topic for stores in West Midlands and South West are greater than the probability estimates of the Welsh topic at stores in further regions. Moreover, the Gaussian process captures spatial residual distinguishing the stores in the neighbouring regions that are close to Wales from the stores (in the same regions) that are at further distances, as demonstrated in Figure 7 (right panel).

The coefficients for the North and Centre topic clearly show that the topic is more likely in the North West, North East, Yorkshire and West Midlands and is less likely in Northern Ireland and Scotland. On the other hand, the coefficients for the South and Midlands show that on average the topic is more likely in the southern and central English regions; however, only the coefficient of East England has a non-zero 95% credibility interval. The Organic topic shows a different pattern, its average coefficients are negative; this indicates that the probability of the Organic topic is on average lower than the average topic probability in London. In other words, the Organic topic is more likely in London than in any other region or constituent country; however, the coefficients show 95% credible intervals containing zero, suggesting that the regional effect may not be significant.

Chapter 3.3 of the 2012 Family Food report [44] explores regional comparisons within England, highlighting several similar regional differences which are model is also able to identify. For example, our CentralSouth topic is very similar to the Central-North, except the latter generally contains higher fat content: bread is replaced by muffins, ham is replaced by bacon, resonating with the North–South differences of the Family Food survey.

Similarly, potatoes featured in the top 15 products of our Central-North topic, but not the CentralSouth. Finally, our Cornish topic showed high cream consumption, in-line with the Food Survey. On the other hand, in the country-level comparison of the Family Food report, Northern Island shows the highest consumption of potatoes, whereas in our analysis, the Welsh topic displays even wider variety of potatoes. This could either indicate a new trend, or reflect the fact that our analysis disregards quantities of items within each transaction.

The covariance parameters length-scale ρ and amplitude α model the covariance between stores, which is stronger when the spatial distance is smaller than ρ and when α is significantly larger from zero. The Welsh and the Organic topic show strong covariance as depicted in Figure 8. On the other hand, the Northern Irish topic and the Scottish topic show small values of α indicating weak covariance functions.

6.2 | Linear Gaussian Process Regression Versus Linear Regression

Here, we compare *mean squared error* and the log of the probability density on held-out data obtained from model topic prevalence using linear GPR and the linear regression (LR). We will show that the former model retrieves more accurate estimates and better predictive likelihood by modelling residual spatial effect.

Table 2 shows that GPR improves the prediction of topic probabilities of the Welsh, English-Northern and Centre, South and Midlands and Organic topics. The difference between the mean

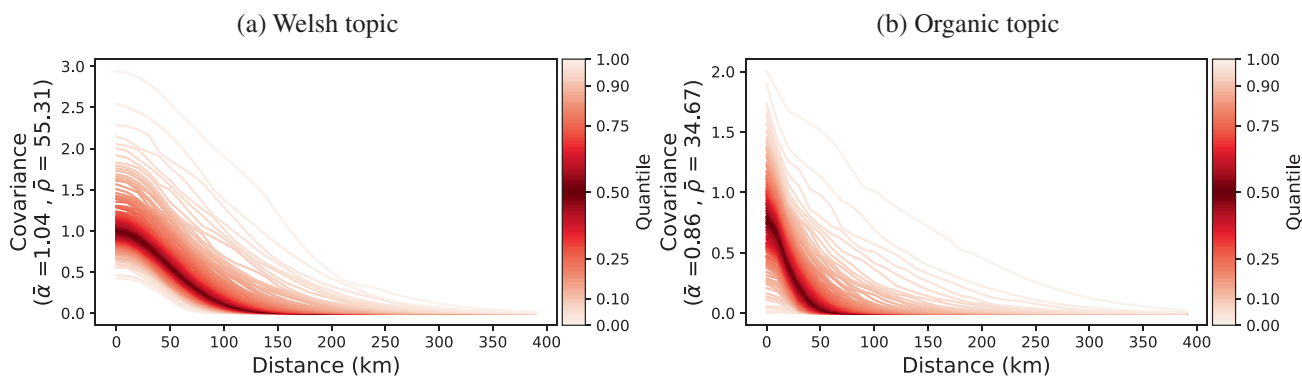


FIGURE 8 | Covariance function of the (a) Welsh topic and (b) Organic topic. Lines are computed with posterior samples of α and ρ .

TABLE 2 | Comparison of the linear Gaussian process regression (GPR) versus linear regression (LR).

	Northern Irish	Scottish	Welsh	English-North and Centre	English-South and Midlands	Organic
LR: MSE (SE)	0.64 (0.001)	2.16 (0.004)	3.18 (0.006)	3.36 (0.007)	1.65 (0.004)	3.39 (0.007)
GPR: MSE (SE)	0.63 (0.001)	2.15 (0.004)	2.64 (0.005)	3.24 (0.004)	1.62 (0.003)	3.18 (0.006)
<i>p</i> -value	0.5877	0.1664	0.0000	0.0000	0.0000	0.0000
LR lppd (SE)	−298.3 (0.30)	−450.2 (0.25)	−499.1 (0.23)	−513.5 (0.31)	−418.8 (0.38)	−506 (0.26)
GPR lppd (SE)	−296.5 (0.28)	−449.3 (0.26)	−476.9 (0.26)	−504.9 (0.43)	−412.6 (0.40)	−493.9 (0.27)
<i>p</i> -value	0.0000	0.0169	0.0000	0.0000	0.0000	0.0000

Note: *p*-values are computed for the pointwise difference of the two methods at each observation in the test set.

Abbreviation: lppd: log posterior predictive density on test data.

squared error of these topics is statistically significant at the 0.05 level, indicating that the Gaussian process provides significant model improvement. Similarly, the log predictive likelihood of the four aforementioned topics is significantly better at the 0.05 level. On the contrary, the GPR doesn't show significantly improved predictions of the Scottish and Northern Irish topics. The difference of their mean squared errors is not statistically significant at the 0.05 level; however, the GPR shows significantly better predictive log-likelihood at the 0.05 level.

Examining GPR residuals in Figure 9, we still observe spatial patterns that are not captured by the Gaussian process. For example, concentrations of underestimated probabilities around North West in Figure 9a, around the centre of Scotland in Figure 9b, around South West and East Anglia in Figure 9c; and overestimated probabilities around South East in Figure 9b. Further work could explore the Gaussian process with non-stationary covariance to capture local spatial patterns.

7 | Discussion

In this article, we showed that STM is powerful in the analysis of transaction retail data, identifying topics that characterise various customer needs, particularly, those that reflect regional demand. STM harnesses store structure, describing transactions and stores as mixtures of objectives. More importantly, STM can identify regional topics that otherwise would be overseen by the widely used topic model, the LDA. Aggregating multiple samples of the posterior distribution and selecting topic modes allow the identification of certain and meaningful topics, achieving better data representations and capturing posterior variability. Topic analysis, through GPR, quantifies regional effects and captures

spatial dependence through the squared exponential covariance function.

7.1 | Computational Considerations

The computational aspect of our approach is the biggest bottleneck of the model. Both the Gibbs sampler for the STM, as well as the Hamiltonian Monte Carlo sampler implemented through R-Stan, become prohibitively slow as sample sizes increases.

The parameter space of LDA and related models is vast and discrete and thus exploration through a Gibbs sampler is slow, even though eventual convergence is guaranteed. Tuning the Gibbs sampler (for example, by using adaptive Gibbs sampling [45]) may be a promising avenue to explore. In our experience, variational approaches using the mean-field approximation [46] do not produce topics of satisfactory quality for these types of data (see discussion in our earlier paper [2]); however, recent advances using amortised inference [47] may offer improvements.

GP regression using a large number of spatial points is also computationally intensive, due to the underlying dense covariance matrix. Computational shortcuts using a basis function approximation via Laplace eigenfunctions have been shown to offer results of similar accuracy at a fraction of the computational cost for stationary GPs [48]. These improvements could also allow the analysis of spatial topics using other geographical hierarchies such as middle layer super output areas.

7.2 | Model Improvement and Future Work

Another consideration of the proposed method is the relationship of the two-step approach to cut posteriors [49]. In this

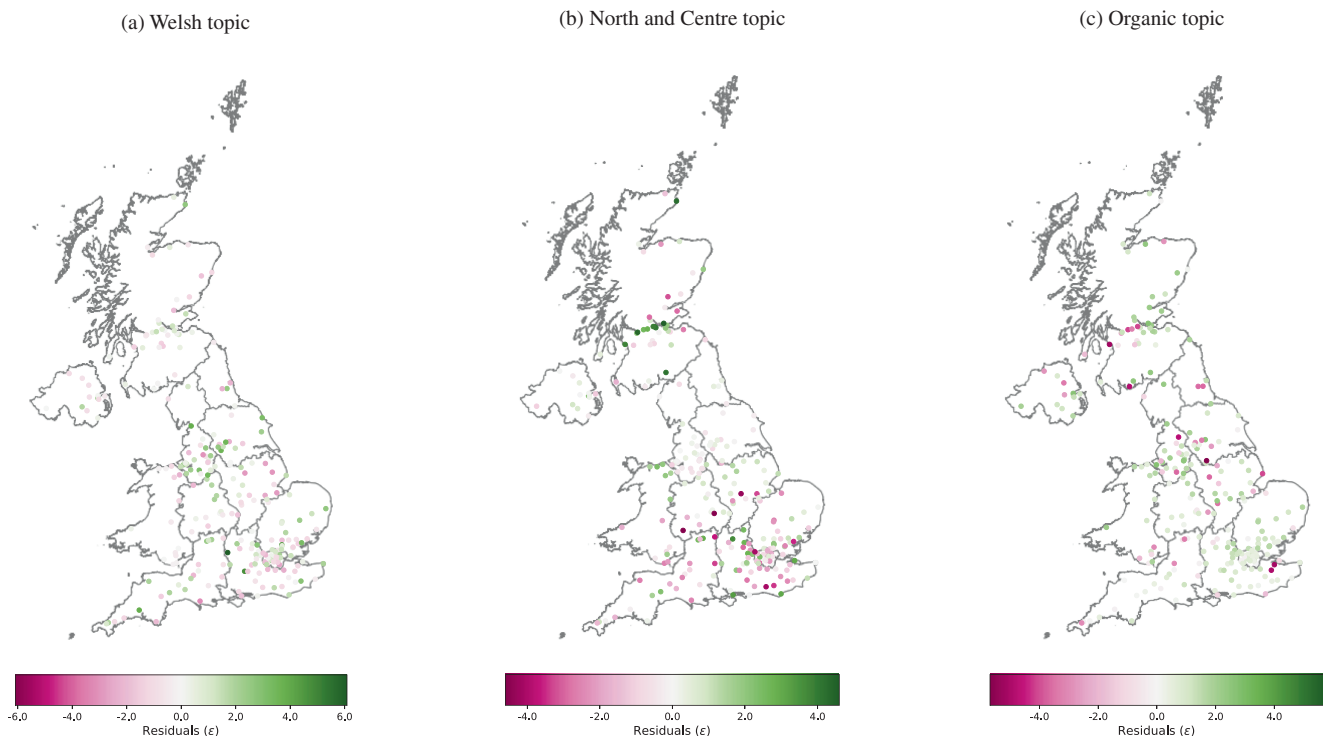


FIGURE 9 | Residuals of modelling the Welsh/North and Centre/Organic topic with GPR.

article, we fit the GP regression on posterior means of the STM sampler. However, Equation (11) could be replaced by a set of t posterior samples rather than a posterior mean estimate; these posterior samples would then form t draws from the GP process, and would then be fitted as independent draws of the entire set of geographical locations. This corresponds to a cut posterior model and would capture not only the mean topic probabilities but also their posterior uncertainty from the STM samplers. We have not pursued this direction in this article due to the computational bottleneck of the GP in R-Stan, but this is part of ongoing work.

A final consideration of model validity is the suitability of topic models given their assumption of independent product draws conditionally on the underlying set of topics. This assumption is not valid in practice, since customers typically shop with some version of a shopping list. For example, a topic including cleaning products might contain two different brands of washing powder with high probability; however, customers would be unlikely to purchase both at the same time. This type of model misfit would be particularly harmful if interest lay in predicting future grocery baskets. In our case, however, ignoring that dependence and focusing directly on marginal product inclusion probabilities still reveals valuable and interpretable output.

Acknowledgments

This work was supported by ESRC Grant No. ESRC ES/P000592/1 and The Alan Turing Institute under the UK EPSRC Grant No. EP/N510129/1.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available from dunnhumby Ltd. Restrictions apply to the availability of these data, which were used under license for this study. Data are available from the author(s) with the permission of dunnhumby Ltd.

References

1. L. Du, W. Buntine, and H. Jin, "A Segmented Topic Model Based on the Two-Parameter Poisson-Dirichlet Process," *Machine Learning* 81, no. 1 (2010): 5–19.
2. M. Vega Carrasco, I. Manolopoulou, J. O'Sullivan, R. Prior, and M. Musolesi, "Posterior Summaries of Grocery Retail Topic Models: Evaluation, Interpretability and Credibility," *Journal of the Royal Statistical Society: Series C: Applied Statistics* 71 (2022): 562–588.
3. S. Banerjee, B. P. Carlin, and A. E. Gelfand, *Hierarchical Modeling and Analysis for Spatial Data* (Boca Raton, FL: CRC Press, 2014).
4. N. Cressie and C. K. Wikle, *Statistics for Spatio-Temporal Data* (Hoboken, NJ: John Wiley & Sons, 2015).
5. C. K. Williams and C. E. Rasmussen, *Gaussian Processes for Machine Learning* (Cambridge, MA: MIT Press, 2006).
6. K. Christidis, D. Apostolou, and G. Mentzas, "Exploring Customer Preferences With Probabilistic Topics Models," in *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases* (Athens, Greece: National Technical University of Athens, 2010), 12–24.
7. H. Hruschka, "Linking Multi-Category Purchases to Latent Activities of Shoppers: Analysing Market Baskets by Topic Models," *Marketing: ZFP – Journal of Research and Management* 36, no. 4 (2014): 267–273.
8. B. J. Jacobs, B. Donkers, and D. Fok, "Model-Based Purchase Predictions for Large Assortments," *Marketing Science* 35, no. 3 (2016): 389–404.
9. H. Hruschka, *Hidden Variable Models for Market Basket Data. Statistical Performance and Managerial Implications*, Technical Report 489 (Regensburg, Germany: University of Regensburg, Department of Economics, 2016).
10. N. Schröder, "Using Multidimensional Item Response Theory Models to Explain Multi-Category Purchases," *Marketing ZFP* 39, no. 2 (2017): 27–37.
11. A. N. Hornsby, T. Evans, P. S. Riefer, R. Prior, and B. C. Love, "Conceptual Organization is Revealed by Consumer Activity Patterns," *Computational Brain & Behavior* 3 (2019): 162–173.
12. C. Sturley, A. Newing, and A. Heppenstall, "Evaluating the Potential of Agent-Based Modelling to Capture Consumer Grocery Retail Store Choice Behaviours," *International Review of Retail, Distribution and Consumer Research* 28, no. 1 (2018): 27–46.
13. A. Davies, L. Dolega, and D. Arribas-Bel, "Buy Online Collect in-Store: Exploring Grocery Click&Collect Using a National Case Study," *International Journal of Retail & Distribution Management* 47 (2019): 278–291.
14. A. Newing, G. P. Clarke, and M. Clarke, "Developing and Applying a Disaggregated Retail Location Model With Extended Retail Demand Estimations," *Geographical Analysis* 47, no. 3 (2015): 219–239.
15. T. B. Waddington, G. P. Clarke, M. Clarke, and A. Newing, "Open all Hours: Spatiotemporal Fluctuations in UK Grocery Store Sales and Catchment Area Demand," *International Review of Retail, Distribution and Consumer Research* 28, no. 1 (2018): 1–26.
16. T. Berry, A. Newing, D. Davies, and K. Branch, "Using Workplace Population Statistics to Understand Retail Store Performance," *International Review of Retail, Distribution and Consumer Research* 26, no. 4 (2016): 375–395.
17. Office for National Statistics, "Living Costs and Food Survey Quality and Methodology Information (QMI)," August 1, 2023.
18. K. G. Smith, P. Scheelbeek, A. Balmford, P. Alexander, and E. E. Garnett, "Discrepancies Between Two Long-Term Dietary Datasets in the United Kingdom (UK)," *Wellcome Open Research* 6 (2021): 350.
19. R. Smith and K. C. Haverson, "The Greggs-Pret Index: A Machine Learning Analysis of Consumer Habits as a Metric for the Socio-Economic North-South Divide in England," arXiv preprint arXiv:2304.00326, 2023.
20. S. Kuznesof, A. Tregear, and A. Moxey, "Regional Foods: A Consumer Perspective," *British Food Journal* 99, no. 6 (1997): 199–206.
21. A. Groves, *The Local and Regional Food Opportunity* (Watford, UK: Institute of Grocery Distribution, 2005).
22. J. Chuang, M. E. Roberts, B. M. Stewart, et al., "TopicCheck: Interactive Alignment for Assessing Topic Model Stability," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Denver, CO: Association for Computational Linguistics, 2015), 175–184.
23. M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, "The Author-Topic Model for Authors and Documents," in *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence* (Arlington, VA: AUAI Press, 2004), 487–494.
24. D. Blei and J. Lafferty, "Correlated Topic Models," in *Proceedings of the 18th International Conference on Neural Information Processing Systems* (Cambridge, MA: MIT Press, 2006), 147–154.
25. M. E. Roberts, B. M. Stewart, D. Tingley, and E. M. Airoldi, "The Structural Topic Model and Applied Social Science," in *Proceedings of*

- the *Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation* (Harrahs and Harveys, Lake Tahoe, 2013), 1–20.
26. D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet Allocation,” *Journal of Machine Learning Research* 3 (2003): 993–1022.
27. D. M. Blei, “Probabilistic Topic Models,” *Communications of the ACM* 55, no. 4 (2012): 77–84.
28. F. Riahi, Z. Zolaktaf, M. Shafiei, and E. Milios, “Finding Expert Users in Community Question Answering,” in *Proceedings of the 21st International Conference on World Wide Web* (New York, NY: ACM, 2012), 791–798.
29. B. Lu, M. Ott, C. Cardie, and B. K. Tsou, “Multi-Aspect Sentiment Analysis With Topic Models,” in *Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops* (New York, NY: IEEE, 2011), 81–88.
30. W. Buntine and M. Hutter, “A Bayesian View of the Poisson-Dirichlet Process,” arXiv preprint arXiv:1007.0296, 2012.
31. H. Ishwaran and L. F. James, “Gibbs Sampling Methods for Stick-Breaking Priors,” *Journal of the American Statistical Association* 96, no. 453 (2001): 161–173.
32. J. Pitman and M. Yor, “The Two-Parameter Poisson-Dirichlet Distribution Derived From a Stable Subordinator,” *Annals of Probability* 25 (1997): 855–900.
33. D. J. Aldous, “Exchangeability and Related Topics,” in *École d’Été de Probabilités de Saint-Flour XIII—1983*, ed. P. L. Hennequin (Berlin, Germany: Springer, 1985), 1–198.
34. C. Chen, L. Du, and W. Buntine, “Sampling Table Configurations for the Hierarchical Poisson-Dirichlet Process,” in *Machine Learning and Knowledge Discovery in Databases*, eds. D. Gunopulos, T. Hofmann, D. Malerba, and M. Vazirgiannis (Berlin, Germany: Springer, 2011), 296–311.
35. G. Malsiner-Walli, S. Frühwirth-Schnatter, and B. Grün, “Model-Based Clustering Based on Sparse Finite Gaussian Mixtures,” *Statistics and Computing* 26, no. 1-2 (2016): 303–324.
36. C. Hennig, M. Meila, F. Murtagh, and R. Rocci, *Handbook of Cluster Analysis* (Boca Raton, FL: CRC Press, 2015).
37. W. R. Tobler, “A Computer Movie Simulating Urban Growth in the Detroit Region,” *Economic Geography* 46, no. sup1 (1970): 234–240.
38. B. Carpenter, A. Gelman, M. D. Hoffman, et al., “Stan: A Probabilistic Programming Language,” *Journal of Statistical Software* 76 (2017): 1.
39. M. Betancourt, “A Conceptual Introduction to Hamiltonian Monte Carlo,” arXiv preprint arXiv:1701.02434, 2017.
40. D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, “Variational Inference: A Review for Statisticians,” *Journal of the American Statistical Association* 112, no. 518 (2017): 859–877.
41. M. D. Hoffman and A. Gelman, “The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo,” *Journal of Machine Learning Research* 15, no. 1 (2014): 1593–1623.
42. Office for National Statistics, accessed September 30, 2019, <http://geoportal1-ons.opendata.arcgis.com/datasets/bf7701adbfc74565a3de7feb414184e8>.
43. C. C. Robusto, “The Cosine-Haversine Formula,” *American Mathematical Monthly* 64, no. 1 (1957): 38–40.
44. Department for Environment, Food and Rural Affairs, *Family Food 2012*, Technical Report (London: Department for Environment, Food and Rural Affairs, 2013).
45. C. Chimisov, K. Latuszynski, and G. Roberts, “Adapting the Gibbs Sampler,” arXiv preprint arXiv:1801.09299, 2018.
46. M. Hoffman, F. Bach, and D. Blei, “Online Learning for Latent Dirichlet Allocation,” in *Proceedings of the 23rd International Conference on Neural Information Processing Systems* (Red Hook, NY: Curran Associates Inc, 2010), 856–864.
47. A. Srivastava and C. Sutton, “Autoencoding Variational Inference for Topic Models,” in *Proceedings of the International Conference on Learning Representations* (2017).
48. G. Riutort-Mayol, P. C. Bürkner, M. R. Andersen, A. Solin, and A. Vehtari, “Practical Hilbert Space Approximate Bayesian Gaussian Processes for Probabilistic Programming,” *Statistics and Computing* 33, no. 1 (2023): 17.
49. D. J. Nott, C. Drovandi, and D. T. Frazier, “Bayesian Inference for Misspecified Generative Models,” *Annual Review of Statistics and Its Application* 11 (2024): 179–202, <https://doi.org/10.1146/annurev-statistics-040522-015915>.

Appendix A

MCMC Convergence Plots

We evaluate four Markov chains of STM with 100 topics. Markov chains are run for 100,000 iterations with a burn-in period of 80,000 iterations. Log-likelihood is measured at every 10 iterations and shown in Figure A1. We calculate the potential scale reduction factor using 8000 samples.

Appendix B

MCMC Convergence of Clustered STM Topics

Once the clustered topics are identified, the STM sampler is re-run with 104 clustered topics fixed and known a priori for 1500 iterations and burn-in period of 1000 iterations. Figure B1 shows convergence of the marginal log-likelihood over MCMC iterations.

Appendix C

Evaluation of Clustered STM Topics

Figure C1a,b shows the evaluation of the clustered topics across cosine distance thresholds.

Appendix D

Block Gibbs Sampler

We use the block Gibbs sampling algorithm proposed in [34] that jointly samples topic assignments and table indicators, leading to a more efficient sampling method. Table counts are not sampled, instead reconstructed by summation of the table indicators:

$$t_k = \sum_{n=1}^N u_n \mathbb{1}_{z_n=k} \quad (\text{D1})$$

Using the table indicator representation, the PDP posterior distribution is:

$$p(z, t|a, b, \theta) = \prod_k \frac{n_k!}{t!(n_k - t)!} p(z, u|a, b, \theta) \quad (\text{D2})$$

responding to $\frac{n_k!}{t!(n_k - t)!}$ sitting arrangements.

The joint distribution of topic assignments and table indicators can be obtained by using Equation (D2) in Equation (2) resulting in:

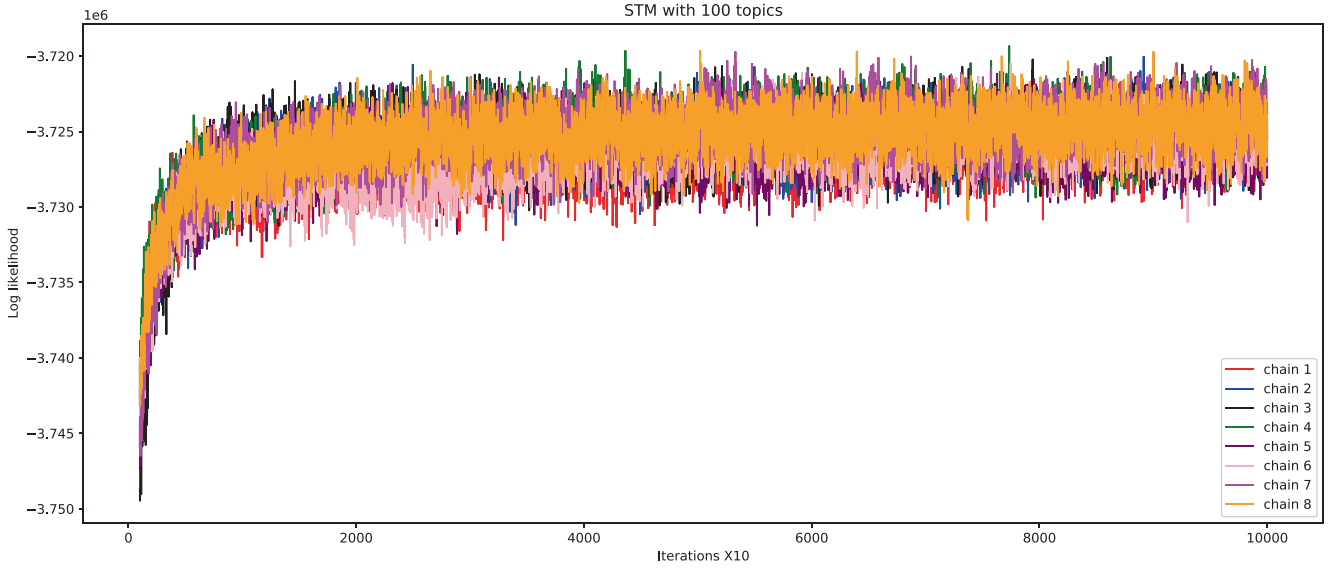


FIGURE A1 | Markov chains of STM with 100. Potential scale reduction factor \hat{R} : 1.07.

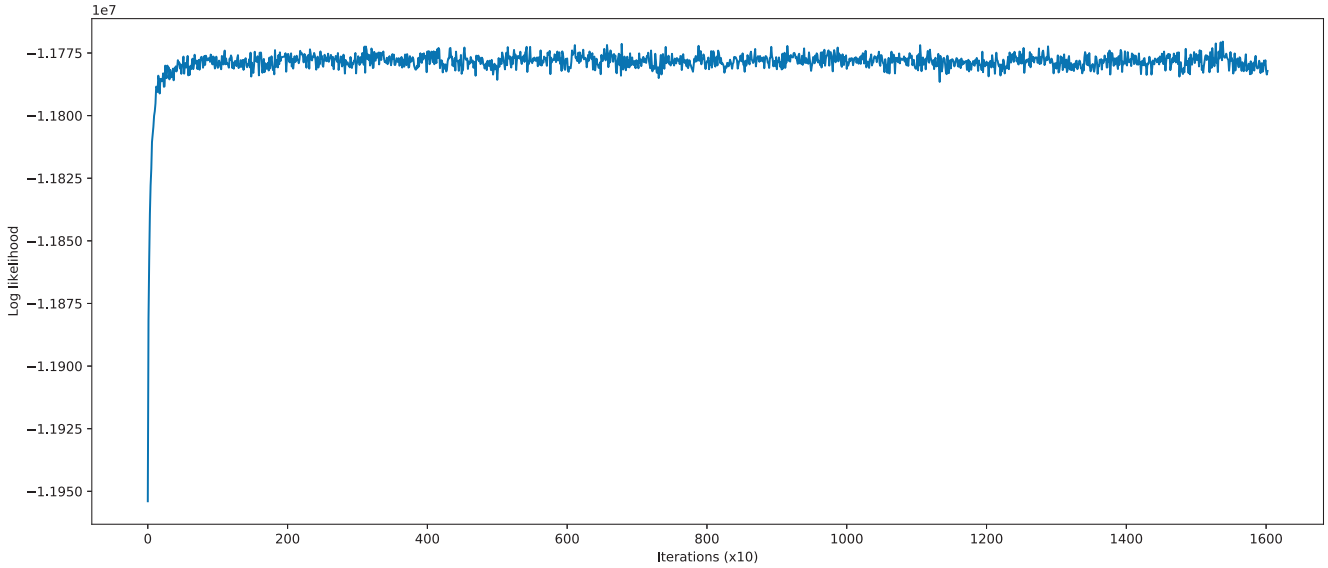


FIGURE B1 | Markov chain of STM with 104 clustered topics known a priori. Potential scale reduction factor \hat{R} : 0.998.

$$\begin{aligned}
 p(\mathbf{z}, \mathbf{w}, \mathbf{t} | \boldsymbol{\alpha}, \boldsymbol{\beta}, a, b) &= \prod_d \frac{\text{Beta}_K(\boldsymbol{\alpha} + \sum_p \mathbf{t}_{p,d})}{\text{Beta}_K(\boldsymbol{\alpha})} \prod_{p,d} \frac{(b|a)_{\sum_k t_{p,d,k}}}{(b)_{N_{p,d}}} \\
 &\prod_{p,d,k} S_{t_{p,d,k}+a}^{N_{k|p,d}} \frac{t_{p,d,k}!(N_{p,d,k} - t_{p,d,k})!}{n_{p,d,k}!} \prod_k \frac{\text{Beta}_V(\boldsymbol{\beta} + \mathbf{N}_k)}{\text{Beta}_V(\boldsymbol{\beta})}
 \end{aligned} \quad (\text{D3})$$

The block Gibbs sampling algorithm first samples a table indicator $u_n = 1$ or $u_n = 0$ with probabilities:

$$p(u_n = 1 | z_n = k) = \frac{t_k}{n_k}, \quad p(u_n = 0 | z_n = k) = 1 - \frac{t_k}{n_k} \quad (\text{D4})$$

and discounts the current assignment z_n from $N_{p,d,k}$ and reduces $t_{p,d,k}$ by 1 if $u_n = 1$.

Then, the full conditional distribution is computed taking into account two scenarios: the probability of opening a new table (Equation D5) and

the probability of choosing an occupied table (Equation D6) if $t'_{p,d,k} > 0$.

$$\begin{aligned}
 p(z_n = k, u_n = 1 | \mathbf{z} - \{z_n\}, \mathbf{u} - \{u_n\}, \mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}, a, b) \\
 \propto \frac{\alpha_k + t'_{d,k}}{\alpha + t'_d} \frac{b + at'_{p,d}}{b + N'_{p,d}} \frac{S_{t'_{p,d,k}+1}^{N'_{p,d,k}+1}}{S_{t'_{p,d,k}}^{N'_{p,d,k}}} \frac{t'_{p,d,k} + 1}{n'_{p,d,k} + 1} \frac{\beta_v + M'_{k,w_{p,d,n}}}{\beta + M'_k}
 \end{aligned} \quad (\text{D5})$$

$$\begin{aligned}
 p(z_n = k, u_n = 0 | \mathbf{z} - \{z_n\}, \mathbf{u} - \{u_n\}, \mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}, a, b) \\
 \propto \frac{S_{t'_{p,d,k}}^{N'_{p,d,k}+1}}{S_{t'_{p,d,k}}^{N'_{p,d,k}}} \frac{1}{b + N'_{p,d}} \frac{n'_{p,d,k} - t'_{p,d,k} + 1}{n'_{p,d,k} + 1} \frac{\beta_v + M'_{k,w_{p,d,n}}}{\beta + M'_k}
 \end{aligned} \quad (\text{D6})$$

where the dash indicates statistics after excluding the current assignment.

Finally, update the counts of $n_{p,d,k}$ and $t_{p,d,k}$ with the sampled topic assignment z_n and table indicator u_n .

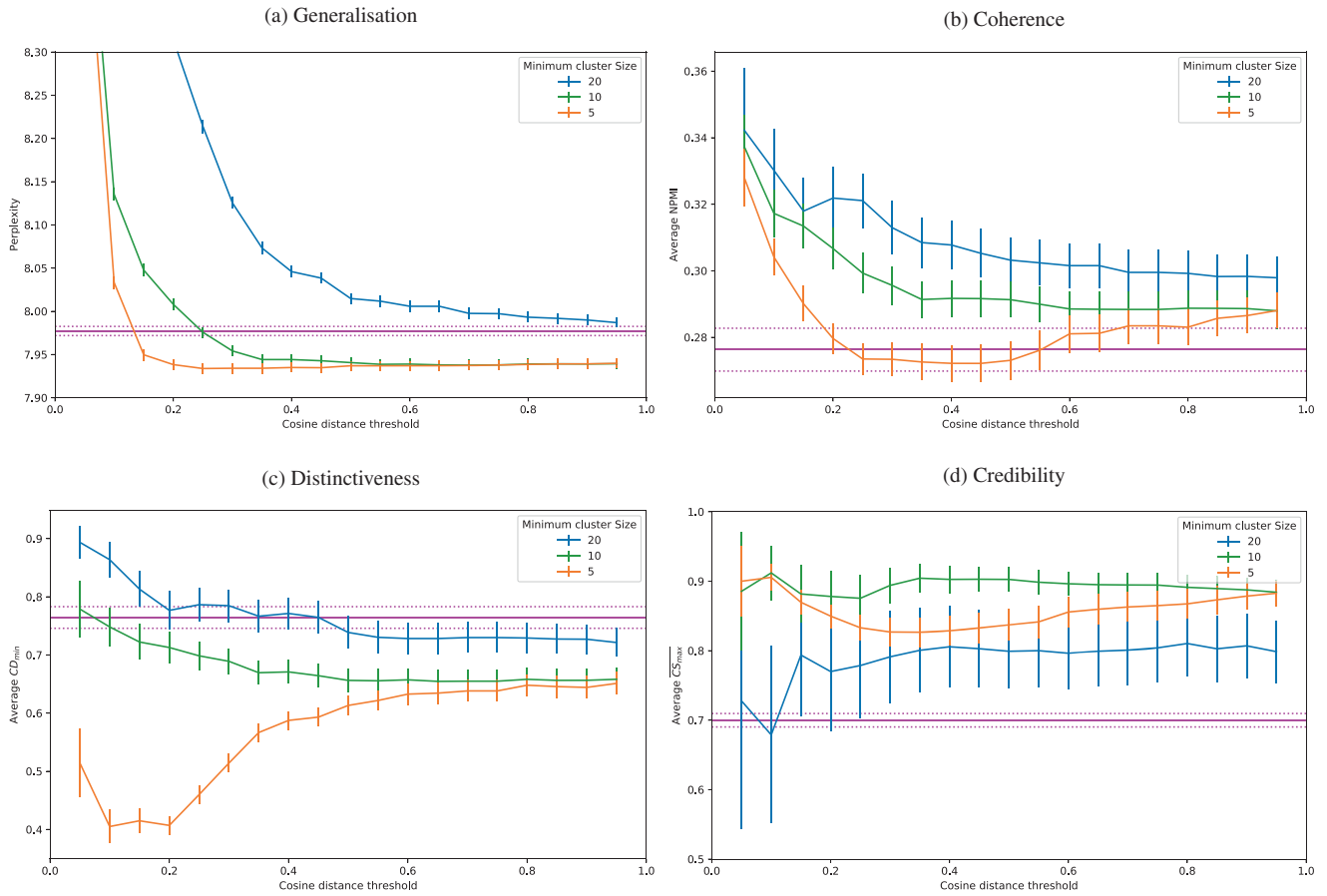


FIGURE C1 | Evaluation of subsets of clustered topics. Subsets are formed with combinations of minimum cluster size and cosine distance thresholds. **Horizontal lines** and **dotted lines** show the average measures (\pm one standard error) of the STM posterior samples. Subsets of clusters formed with a minimum cluster size of 10 show greater coherence and credibility, and the subsets formed with a cosine distance threshold larger than 0.3 show better generalisation (in comparison to the average generalisation of the STM posterior samples). Subsets with a minimum cluster size of 10 show less distinctive clustered topics, which might result from filtering out distinctive but uncertain topics. A cosine distance threshold larger than 0.35 cosine distance does not significantly improve perplexity.

Appendix E

Hierarchical Clustering

The hierarchical clustering algorithm takes a bag of topics, a list with sample indexes, and a cosine distance threshold. The bag of topics gathers topic distributions from various posterior samples from various MCMC. The list of sample indices records a sample index for each topic, that is, assuming that the first 50 topics in the bag of topics come from posterior sample 1 and the next 50 topics come from posterior sample 2, then the first 50 elements in the list of samples indices are 1 and the next 50 elements are 2. The cosine distance threshold indicates the limit up to which topics would be merged.

The algorithm will start by forming clusters with each of the topics in the bag of topics. So, if there are N topics, there are N initial clusters. Then, a list L is created to record the cosine distance between two clusters. This list contains the indexes of the two compared clusters and the cosine distance between the clustered topics. A clustered topic is the average topic distributions of the cluster members.

At each step, the algorithm finds the pair of clusters in L with the minimum cosine distance. Then, the algorithm evaluates if the members

of both clusters are from different posterior samples using the list of sample indices. If so, a new cluster is created by merging the evaluated pair of clusters. Then, the algorithm removes from the L all comparisons that had any of the identified clusters and adds comparisons from all the remaining clusters to the new cluster. But, If the evaluation is false, the algorithm updates the cosine distance between the pair of clusters with 1. Thereby, the algorithm would not take the same pair of clusters in the next step.

The algorithm will keep merging clusters until the minimum cosine distance is larger than the cosine distance threshold. The algorithm then retrieves all the remaining clusters (clusters that are not eliminated because they do not get merged).

Appendix F

Further Topics

Finally, we show the spatial distribution of three more grocery topics in Figure F1.

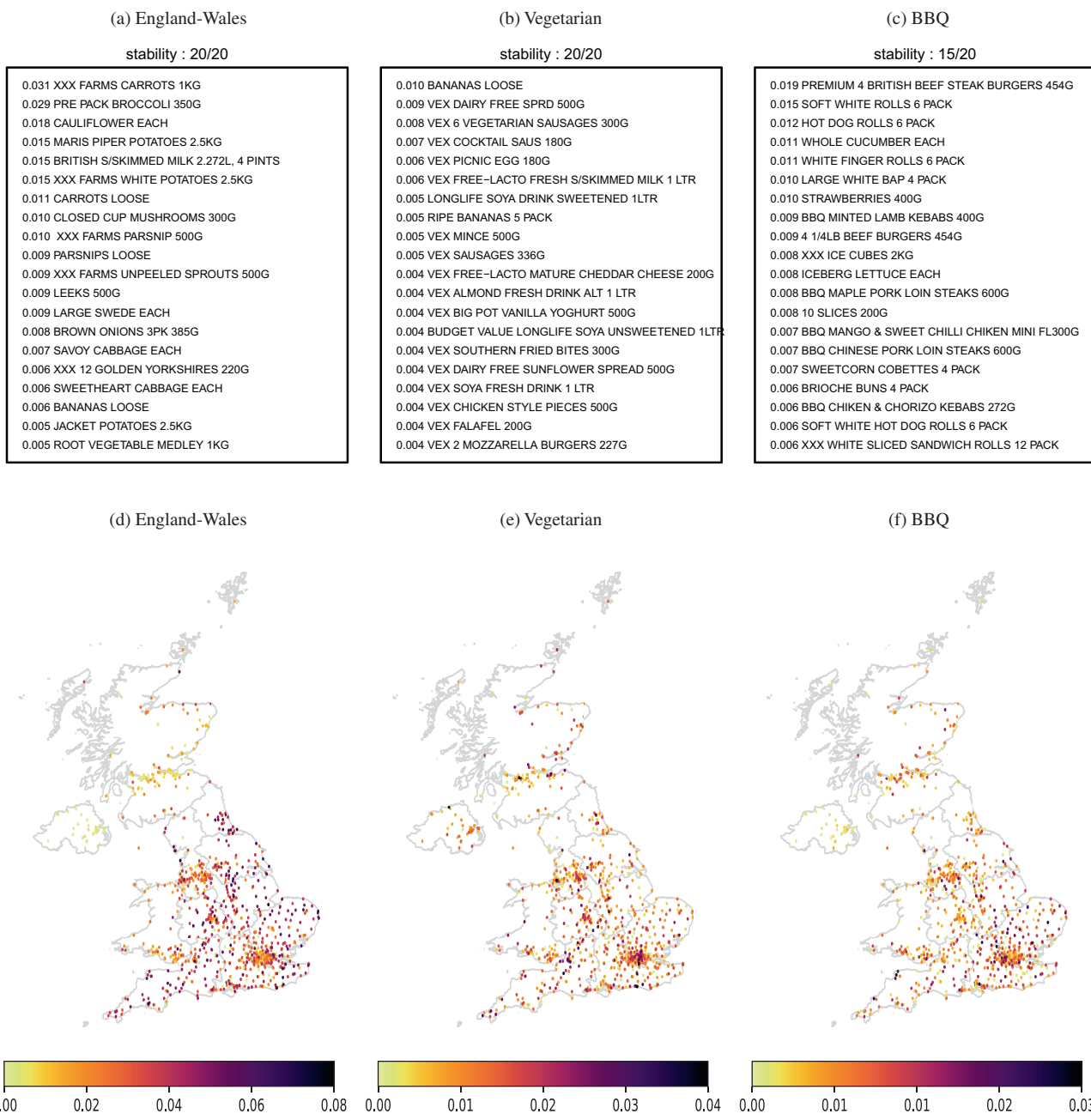


FIGURE F1 | Topic probabilities of clustered grocery topics in the UK for three further topics: (a) England-Wales, (b) Vegetarian and (c) BBQ. The top row shows the top 20 products within each topic. In the bottom row, purple and yellow points reflect the largest and smallest topic probabilities, respectively.