Contents lists available at ScienceDirect

International Journal of Food Microbiology

journal homepage: www.elsevier.com/locate/ijfoodmicro

Probabilistic topic modelling in food spoilage analysis: A case study with Atlantic salmon (*Salmo salar*)

L. Kuuliala[a,b,*], R. Pérez-Fernández[b,1], M. Tang[b], M. Vanderroost[a], B. De Baets[b], F. Devlieghere[a]

[a] Research Unit Food Microbiology and Food Preservation (FMFP), Department of Food Technology, Safety and Health, Part of Food2Know, Faculty of Bioscience Engineering, Ghent University, Coupure links 653, B-9000 Ghent, Belgium

[b] Research Unit Knowledge-based Systems (KERMIT), Department of Data Analysis and Mathematical Modelling, Part of Food2Know, Faculty of Bioscience Engineering, Ghent University, Coupure links 653, B-9000 Ghent, Belgium

ARTICLE INFO

Keywords:
Latent Dirichlet Allocation
Food quality
Metabolomics
Potential spoilage indicator
Volatile organic compound

ABSTRACT

Probabilistic topic modelling is frequently used in machine learning and statistical analysis for extracting latent information from complex datasets. Despite being closely associated with natural language processing and text mining, these methods possess several properties that make them particularly attractive in metabolomics applications where the applicability of traditional multivariate statistics tends to be limited. The aim of the study was thus to introduce probabilistic topic modelling – more specifically, Latent Dirichlet Allocation (LDA) – in a novel experimental context: volatilome-based (sea) food spoilage characterization. This was realized as a case study, focusing on modelling the spoilage of Atlantic salmon (*Salmo salar*) at 4 °C under different gaseous atmospheres (% $CO_2/O_2/N_2$): 0/0/100 (A), air (B), 60/0/40 (C) or 60/40/0 (D). First, an exploratory analysis was performed to optimize the model tunings and to consequently model salmon spoilage under 100% N_2 (A). Based on the obtained results, a systematic spoilage characterization protocol was established and used for identifying potential volatile spoilage indicators under all tested storage conditions. In conclusion, LDA could be used for extracting sets of underlying VOC profiles and identifying those signifying salmon spoilage, giving rise to an extensive discussion regarding the key points associated with model tuning and/or spoilage analysis. The identified compounds were well in accordance with a previously established approach based on partial least squares regression analysis (PLS). Overall, the outcomes of the study not only reflect the promising potential of LDA in spoilage characterization, but also provide several new insights into the development of data-driven methods for food quality analysis.

1. Introduction

The rapid development of metabolomics technologies has greatly improved our understanding of complex biological systems during the past few decades. In the food and nutrition sector, this global trend has had a major impact on the development of foodomics (Miguel et al., 2012), a novel interdisciplinary field where metabolomics – the study of low molecular weight (<1500 Da) molecules associated with biological samples (Castro-Puyana et al., 2017; Pinu, 2016) – has already been used for addressing various questions related to food quality and safety (Böhme et al., 2019; Klampfl, 2018; Mancano et al., 2018; Martinović et al., 2018; Xu, 2017). In particular, the latest advances in the analysis of spoilage-indicating volatile organic compounds (VOCs) have greatly benefitted both scientific knowledge (Dong et al., 2019; Odeyemi et al., 2018; Wang et al., 2016) and technology development (Ghasemi-Varnamkhasti et al., 2018; Pavase et al., 2018; Poghossian et al., 2019).

However, the complexity of the microbial metabolism poses a major challenge in food quality characterization. At any given moment during storage time, the food volatilome consists of numerous compounds that differ in terms of quantity, chemical composition, reactivity, olfactory impact and sensory acceptability. Irrespective of the applied quantification method, the extraction of information from the resulting datasets

* Corresponding author at: Research Unit Food Microbiology and Food Preservation (FMFP), Department of Food Technology, Safety and Health, Faculty of Bioscience Engineering, Ghent University, Coupure links 653, B-9000 Ghent, Belgium.
E-mail address: Lotta.Kuuliala@UGent.be (L. Kuuliala).

[1] Present address: Department of Statistics and O.R. and Mathematics Didactics, University of Oviedo, Calle Federico García Lorca 18, 33007 Oviedo, Spain.

https://doi.org/10.1016/j.ijfoodmicro.2020.108955
Received 1 May 2020; Received in revised form 10 August 2020; Accepted 25 October 2020
Available online 4 November 2020

thus calls for advanced statistical analysis. Basic multivariate methods such as principal components analysis (PCA), partial least squares regression analysis (PLS) and hierarchical cluster analysis (HCA) have frequently been used for this purpose (Bermejo-Prada et al., 2015; Mansur et al., 2019; Mikš-Krajnik et al., 2016). However, the applicability of these methods in biomarker identification tends to be limited; for example, while PLS outperforms HCA and PCA as a selective tool, it still requires a linear relationship between the studied variables and has a limited capacity in distinguishing correlation from a cause-and-effect relationship (Kuuliala et al., 2018). Hence, more flexible methods are needed for improving our ability to identify the most useful volatile spoilage indicators.

Probabilistic topic modelling comprises a group of methods used in machine learning and statistical analysis for extracting underlying thematic information from an unstructured collection of documents (Blei, 2012). Currently, Latent Dirichlet Allocation (LDA) introduced by Blei et al. (2003) represents one of the most widespread approaches. Despite the fact that these methods have traditionally been closely associated with the analysis of textual data – for example, consumer/customer feedback (Bastani et al., 2019; Hu et al., 2019), social media content (Curiskis et al., 2020; Nolasco and Oliveira, 2019) or research interests (Xiong et al., 2019; Yang et al., 2019) – LDA has also already been successfully used in different biological settings, particularly in genomics (Chen et al., 2010; Perina et al., 2010; Pratanwanich and Lio, 2014; Shiraishi et al., 2015; Yu et al., 2014; Zhang et al., 2012). However, to the best of the authors' knowledge, its prospects within food science still remain to be elucidated.

The aim of the present study is to introduce LDA as an exploratory and selective statistical technique for characterizing (sea)food quality on the basis of its volatilome. First, a non-technical overview of the principles of LDA is given in Section 2. In the experimental part (Sections 3–4), LDA is applied for modelling the quality decay of raw Atlantic salmon (*Salmo salar*) under different gaseous atmospheres and for consequently identifying potential spoilage indicators. Special emphasis is given on 1) optimizing the model parameters, 2) developing a systematic spoilage characterization protocol, including a set of criteria for identifying VOCs that hold promising potential for quality monitoring applications (referred to as "potential spoilage indicators" from here on), and 3) comparing the performance of the said protocol with a previously established PLS-based approach (Kuuliala et al., 2019). The obtained results, conclusions and decisions are discussed in Section 5. Finally, summarizing remarks are given in Section 6.

2. Latent Dirichlet Allocation

LDA is a flexible *generative probabilistic model* for discrete data (Blei et al., 2003), meaning that it can be used for determining the joint probability distribution underlying a group of known samples and consecutively generating new samples from the same distribution. This approach not only allows for exploring previously unknown (underlying or *latent*) structures in large and complex datasets, but also reducing dimensionality, detecting co-occurring variables and evaluating the similarity between individual unlabeled samples. It should be noted though that since LDA is inherently *unsupervised*, it does not involve a pre-defined output and thus cannot be directly used for analyzing the relations between independent and dependent variables. For that purpose, a complementary modelling approach and/or an extension into (semi-)supervised LDA is needed (for further information, see e.g. Fu et al., 2015; Li et al., 2018).

The key concepts of topic modelling are *word* (*term*), *document, corpus*, *word-document matrix (WDM)* and *topic*. By definition, a word refers to a basic unit of discrete data, a document to a sequence of words, and a corpus to a collection of documents (Blei et al., 2003). A WDM (also known as a *document-term matrix* or a *bag of words*) indicates the frequency of each word in each document belonging to the corpus; in case of n documents and m words, an $n \times m$ WDM is obtained (Liu et al., 2016). This information can be used for extracting a set of probability distributions over all words which appear in the corpus (i.e. the *vocabulary*); these distributions – or, more specifically, the interpretations of these distributions (see Section 3.2.3) – can be referred to as topics. In other words, a given document is seen as a product of a generative process, consisting of 1) choosing a distribution over topics, 2) a topic from the chosen distribution, 3) a word from the chosen topic, and 4) returning to step 2 until the pre-determined document length has been reached (Blei et al., 2003).

The application of LDA in a classical text mining setting is illustrated in Example 1.

Example 1. A corpus was formed from 121 abstracts (documents) of original research conducted at the Research Unit Food Microbiology and Food Preservation at Ghent University (FMFP) and published in international peer-reviewed journals between 2000 and 2018. A WDM was constructed by calculating the frequency of each individual word in each abstract (excluding digits, punctuation, symbols and English stopwords) and used for generating an LDA model with five topics. The R package **tm** (Feinerer and Hornik, 2017; Feinerer et al., 2008) was used for constructing the WDM and the package **topicmodels** (Grün and Hornik, 2011) for learning the LDA model with default parameters. The obtained results were examined and visualized in accordance with Section 3.2.3.

Fig. 1 presents the 20 most common words and their distribution in the extracted topics 1–5, associatively interpreted as follows: 1) microbial aspects of food preservation technologies, 2) food quality and microbial spoilage, 3) modelling of microbial behavior in foods, with special emphasis on norovirus, 4) food packaging and shelf life, 5) microbial food safety and health. Overall, these topics could be interpreted as the central themes of research carried out at FMFP during the studied time interval. It should be noted that the number of topics affects their specificity; for example, a model with two topics could be interpreted as 1) food quality and 2) food safety (data not shown).

Table 1 shows the distribution of topics 1–5 in selected documents. For example, document 16 ("Growth of Escherichia coli O157:H7 and Listeria monocytogenes with prior resistance to intense pulsed light and lactic acid" by Rajkovic et al., 2011) could be associated with food preservation (topic 1), whereas the 64:36 relation between topics 1 and 3 in document 40 ("Multi-method approach indicates no presence of sub-lethally injured Listeria monocytogenes cells after mild heat treatment" by Uyttendaele et al., 2008) indicates that pathogen growth was examined by both experimental and statistical methods.

When compared with classical multivariate methods, LDA poses several advantages that make it particularly attractive for addressing complex biological problems. It is flexible, adaptable and imposes relatively few assumptions; importantly, the number of topics (k) is assumed to be known, the order of words within a given document to be exchangeable and all documents to be independent of each other (Liu et al., 2016). A given document may be associated with multiple topics and a given word may belong to multiple documents (Binkley et al., 2014; Griffiths and Steyvers, 2004). Overall, LDA is compatible with any inference algorithm and can be extended or incorporated as part of a more complex approach (Blei et al., 2003). A further discussion on model tuning and associated challenges is provided in Section 5.1.

3. Materials and methods

3.1. Data pre-processing and notation

All statistical analyses were performed using R 3.6.1 (R Core Team, 2019) on four independent subsets (A-D) of the salmon data previously collected by Kuuliala et al. (2019). Briefly, the experimental units of the study were individually packed salmon fillet portions, stored for ≤13 days at 4 °C under specific gaseous atmospheres (% $CO_2/O_2/N_2$): 0/0/100 (A), air (B), 60/0/40 (C) or 60/40/0 (D). Each package ($n = 16$ per

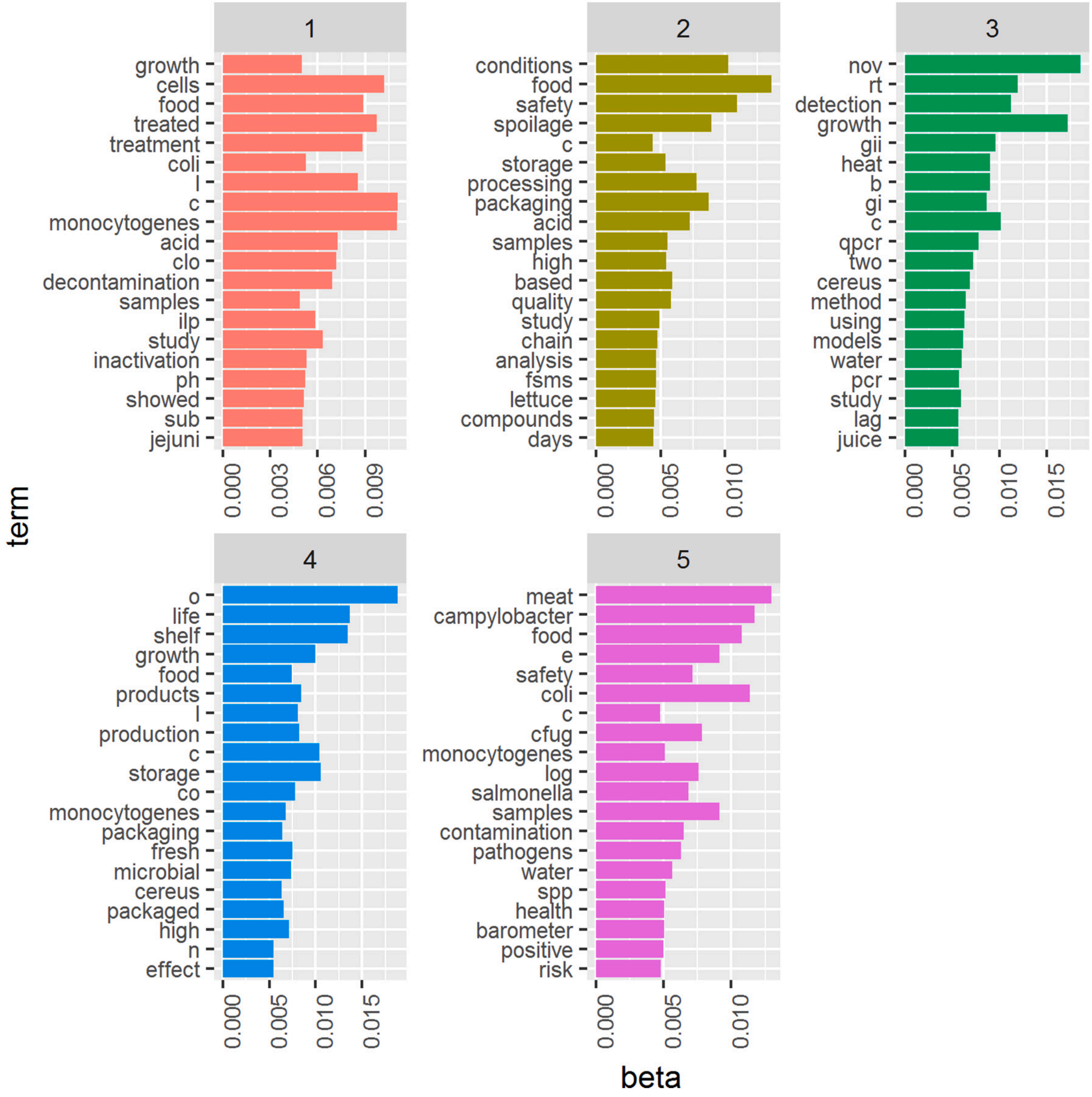


Fig. 1. The distribution of the top-20 words in five topics extracted from a collection of 121 original research abstracts, published by the Research Unit Food Microbiology and Food Preservation (FMFP; Ghent University, Ghent Belgium) between 2000 and 2018.

atmosphere) was linked with a single value of each of the following variables: storage time (d), concentrations of 25 VOCs (C1-C25: ppb) and sensory rejection percentage ($R_{\%}$: %). These subsets were used for generating four WDMs, where the individual salmon packages were treated as “documents”, VOCs as “words” and their concentrations rounded to the nearest whole numbers as “frequencies”. Following the same semantic principles, the extracted “topics” are referred to as “VOC profiles” or “profiles” throughout the present manuscript and denoted whenever applicable as X*k*P*n*, where X indicates the subset (A-D), *k* the number of profiles, and *n* is a profile identifier ($n = 1,2, \ldots,k$).

3.2. Exploratory analysis

The exploratory analysis aimed at optimizing model performance (Sections 3.2.1–3.2.2) and, consecutively, establishing the principles of spoilage characterization (Sections 3.2.3–3.2.4). All activities were performed using the subset A (100% N_2) and its WDM. In case of all mentioned R functions, default parameters were used unless otherwise specified.

3.2.1. Model tuning

In literature, several metrics have been proposed for facilitating the selection of the number of topics (here, profiles). However, the existing methods typically apply different identification criteria; for example, the algorithm of Cao et al. (2009) returns the value of *k* that minimizes the average cosine distance between the extracted topics, the algorithm of Arun et al. (2010) the value that minimizes the symmetric Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951) between the matrices representing the word-per-topic and topic-per-document distributions, and the algorithm of Deveaud et al. (2014) the value that maximizes the sum of the divergences between topic pairs. For this reason, multiple metrics are frequently implemented for comparative purposes.

In this study, the three aforementioned metrics (denoted Cao, Arun and Deveaud) were used for selecting appropriate tunings for LDA models with 2–10 profiles, using the function FindTopicsNumber() from the package **ldatuning** (Murzintcev, 2019). Inference was estimated by the variational expectation-maximization algorithm (VEM) or Gibbs sampling with following specifications:

- VEM: *method*=”VEM”

Table 1
The distribution (%) of topics T1 - T5 in every 8th document in a collection of 121 original research abstracts, published by the Research Unit Food Microbiology and Food Preservation (FMFP; Ghent University, Ghent Belgium) between 2000 and 2018.

ID	Bibliographic reference	T1	T2	T3	T4	T5
8	Zhang et al., 2014. LWT Food Sci. Tech. 55, 224–231.	0.03	0.03	0.03	99.9	0.03
16	Rajkovic et al., 2011. Food Microbiol. 28 (2011), 869–872.	99.92	0.02	0.02	0.02	0.02
24	Peelman et al.. 2014. Innov. Food Sci. Em. Tech. 26, 319–329.	0.01	0.01	0.01	99.94	0.01
32	Tong Thi et al., 2015. Int. J. Food Microbiol. 208, 93–101.	26.79	0.02	0.02	73.17	0.02
40	Uyttendaele et al., 2008. Int. J. Food Microbiol. 123, 262–268.	64.23	0.02	35.71	0.02	0.02
48	Devlieghere et al., 2009. Int. J. Food Sci. Tech. 44, 337–341.	0.04	0.04	0.04	27.81	72.07
56	Daelman et al., 2013. Int. J. Food Microbiol. 160, 193–200.	0.01	0.01	0.01	99.96	0.01
64	Li et al. 2018. Food Control 89, 235–240.	0.02	0.02	0.02	0.02	99.93
72	Li et al., 2011. J. Virol. Methods 177, 153–159.	68.6	0.02	31.35	0.02	0.02
80	Li et al., 2012. J. Virol. Methods 181, 1–5.	54.43	0.02	45.5	0.02	0.02
88	Habib et al., 2012. Food Microbiol. 29, 105–112.	0.01	0.01	0.01	0.01	99.95
96	Wilmart et al., 2015. Eur. J. Plant Pathol. 141, 349–360.	0.02	0.02	0.02	0.02	99.92
104	Jacxsens et al., 2001. Int. J. Food Microbiol. 71, 197–210.	0.01	0.01	0.01	99.94	0.01
112	Baert et al., 2008. Int. J. Food Microbiol. 123, 101–108.	0.01	0.01	99.94	0.01	0.01
120	Rajkovic et al., 2006. Int. J. Food Sci. Technol. 41, 878–884.	0.02	0.02	0.02	99.93	0.02

- Gibbs100A: *method*="Gibbs", $iter = 100$, $burnin = 0$, $thin = 1$
- Gibbs100B: *method*="Gibbs", $iter = 100$, $burnin = 50$, $thin = 1$
- Gibbs1000A: *method*="Gibbs", $iter = 1000$, $burnin = 0$, $thin = 1$
- Gibbs1000B: *method*="Gibbs", $iter = 1000$, $burnin = 500$, $thin = 1$

where *iter* refers to the number of iterations, *burnin* to the number of discarded (burned) initial iterations and *thin* to the interval of retained iterations. For evaluating modelling consistency, two lists consisting of ten random seeds (s1-s10) or WDM column orders (O1-O10) were created; all possible seed-order combinations were tested using each of the aforementioned five methods. Finally, the obtained results (optimal number of profiles: k_{opt}) were visualized using the function heatmap.2() from the package **gplots** (Warnes et al., 2020). In conclusion (see Section 5.1 for a discussion), the method Gibbs1000B and a single seed-order pair were selected to be used for all further modelling activities.

3.2.2. Cross-validation

In literature, the performance of LDA is frequently assessed by means of a perplexity analysis. Briefly, perplexity ($P(w)$) is a measure of the model's ability to predict an unseen dataset (w) and can be denoted as follows:

$$P(w) = exp\left(-\frac{log(p(w))}{\sum_{d=1}^{D}\sum_{j=1}^{V} n^{jd}} \right) \tag{1}$$

where log ($p(w)$) is the log-likelihood of the new data and n^{jd} the number of times the j^{th} term occurs in the d^{th} document (Grün and Hornik, 2011); an increasing perplexity signifies a decreasing model performance.

In this study, perplexity analysis was used for specifying the selection of k and for evaluating the prediction ability. First, leave-one-out cross-validation was performed with functions provided in the package **topicmodels** (Grün and Hornik, 2011). Each sample 1–16 was assigned once as the holdout set while using the other fifteen samples as the training set; models ($k = 2, \ldots,10$) were learned using the function LDA() and the selected tunings (Section 3.2.1). Perplexities were calculated for each model with the function perplexity(), using the holdout sets as the new data.

3.2.3. Profile interpretation

In literature, different methods are used for interpreting the extracted LDA profiles. In classical text mining applications, this task is often carried out by examining the most frequent words of each extracted topic (see, e.g., Example 1). Although it may seem an intuitive process, it relies on knowledge regarding the logical relations between different words: semantically speaking, the concept of "topic" should thus be seen as the "label" of a word distribution rather than being synonymous with the distribution itself.

In this study, profile interpretation was performed for identifying spoilage-associated profiles. First, three exploratory models ($k = 3, 5, 9$) were learned using the function LDA() from the package **topicmodels** and the selected tunings (Section 3.2.1). Relevant distributions were extracted using the function tidy() from the package **tidytext** (Silge and Robinson, 2016): the distribution of the VOCs within profiles was visualized using the package **ggplot2** (Wickham, 2016) and the distribution of the profiles within samples using Excel 2013 for Windows. In order to reduce the risk of interpretation bias, top-8 VOCs were reported in accordance with the suggestion of Agrawal et al. (2018). For assessing the similarity between the compositions of all extracted profiles, hierarchical cluster analysis (HCA) was performed on the basis of the Euclidean distance and average linkage, using the function pvclust() from the package **pvclust** (Suzuki et al., 2019). Finally, the profiles were interpreted by examining the relations between cluster distribution, storage time and $R_{\%}$ (Excel 2013 for Windows).

3.2.4. Identification of potential spoilage indicators

In this study, potential spoilage indicators were identified by evaluating the prevalence of each VOC in the spoilage-associated profiles of the three exploratory models (Section 3.2.3). This was done by establishing the relative spoilage association ($SA_{\%}$):

$$SA_{\%} = \frac{n_{sc}}{n_{tot}} \tag{2}$$

where n_{sc} is the number of times a given compound occurred among the top-8 VOCs of the spoilage-associated profile(s) ($n = n_{tot}$) of a given model ($k = 3, 5, 9$). Compounds fulfilling the criterion $SA_{\%} \geq 0.5$ (occurrence among the top-8 VOCs in at least half of the spoilage-associated profiles) were considered relevant. Finally, identification performance was assessed by comparing the obtained results with the outcomes of a previously established PLS-based protocol (Kuuliala et al., 2019), denoted here as IP_{PLS}.

3.3. Selective analysis

The selective analysis aimed at the identification of potential spoilage indicators under all tested storage conditions. This was done by applying the spoilage characterization protocol developed during exploratory analysis (Section 4.1.5) also for subsets B-D. Briefly, independent LDA models ($k = 3$) representing a given condition (B-D) were learned and interpreted in accordance with Section 3.2.3 and potential spoilage indicators were identified in accordance with Section 3.2.4.

4. Results

4.1. Exploratory analysis

4.1.1. Model tuning

The impact of model tuning on the optimal number of profiles (k_{opt}) is visualized in Fig. 2. The Cao and Deveaud metrics were found to give highly similar results, the median (Md) being either four (VEM-Cao, VEM-Deveaud, Gibbs100A-Deveaud, Gibbs100B-Deveaud), five (Gibbs100B-Cao, all Gibbs1000-methods) or six (Gibbs100A-Cao). In contrast, the Arun metric differed from the other two by consistently suggesting Md(k_{opt}) = 2, irrespective of the chosen tunings. The smallest observed variation in k_{opt} and thus the best model stability was achieved with the method Gibbs1000B; generally, increasing the iteration number (100 vs. 1000) and/or the number of burned iterations (0 vs. 50 or 500) reduced the variation in case of the Cao metric, whereas respective stabilization in the Deveaud metric was achieved after a parallel increase in both parameters (100A vs. 1000B). Finally, no clear trends associated with WDM column order and/or seed could be detected under any tested circumstances.

4.1.2. Cross-validation

The perplexities of the cross-validation models are shown in Table 2. When considering any given holdout sample (ID = 1, …,16), little difference could typically be observed between different levels of $k \neq 2$. On the other hand, when considering any given k, a distinct pattern could be observed over storage time. The highest values (7.60–52.30) appeared in the beginning of storage time (days 1–3), followed by a decrease between days 5 and 13 (1.78–3.63). The prediction ability was thus lowest in the case of samples analyzed during the early days of storage.

4.1.3. Profile interpretation

The distributions of the top-8 VOCs within the extracted profiles (A3P1-P3, A5P1-P5, A9P1-P9) of the three exploratory models are shown in Fig. 3A-C and the corresponding clustering results in Fig. 3D. Irrespective of the value of k, three main profile clusters could be observed. In cluster 1 (A3P1, A5P1 and A9P5), none of the top-8 VOCs (ethanol, 3-methyl-1-butanol, dimethyl sulfide, carbon disulfide, acetone, ammonia, 2,3-butanedione, 3-methylbutanal, acetic acid and/or ethyl acetate) accounted for over 25% (in terms of relative abundance) of the entire volatilome, whereas cluster 2 (A3P3, A5P2, A5P5, A9P3, A9P4 and A9P8) was dominated by hydrogen sulfide + ethanol and cluster 3 (A3P2, A5P3, A5P4, A9P1, A9P2, A9P6, A9P7 and A9P9) by ethanol. In the latter two cases, the most abundant compound(s) accounted for over 80% of the extracted profiles.

The relations between profile distribution, storage time and sensory rejection are shown in Fig. 4. When considering storage time (Fig. 4A-C-E), a major shift in volatilome composition could be observed between days 3–5 in all three models. Cluster 1 was found most prominent during the early days (1–3) and cluster 3 during the latter days (5–11), whereas cluster 2 had a minor contribution and only exceeded 50% in two late-stage samples (days 11–13). When considering sensory rejection (Fig. 4B-D-F), respective observations could be made; generally, increasing $R_{\%}$ was accompanied with decreasing contribution of cluster 1 and corresponding increase in clusters 2–3. Cluster 1 dominated the volatilome of samples with less than 10% rejection, whereas clusters 2 and 3 had varying contributions at a certain rejection percentage. Finally, when comparing the profile distributions of any given sample (Fig. 4A-C-E), increasing k was seen to cause a partitioning in subprofiles while retaining the three main clusters. For example, the partitioning of A3P2 ($k = 3$) into A5P4 and A5P3 ($k = 5$) and further to A9P1, A9P2, A9P6, A9P7 and A9P9 ($k = 9$) could be observed under cluster 2. In conclusion, profiles belonging to clusters 2 and 3 were considered spoilage-associated (see Section 5.2 for discussion) and were thus used in $SA_{\%}$ calculations (Section 4.1.4).

4.1.4. Identification of potential spoilage indicators

The relative spoilage associations of all quantified VOCs are given in Table 3. A comparison between the models with 3, 5 or 9 profiles showed that 11, 9 and 5 compounds had $SA_{\%} \geq 0.5$, respectively. More specifically, three major VOC groups could be identified: 1) 13/25 compounds with $SA_{\%} < 0.5$ in all three models (C2, C7-C10, C13, C15, C17, C18, C22-C25), 2) 5/25 compounds showing an initial $SA_{\%} = 1$ and a decreasing trend along with increasing k (C4, C16, C19–21) and 3) 7/25 compounds fluctuating around $SA_{\%} = 0.5$ (C1, C3, C5, C6, C11, C12, C14).

4.1.5. Spoilage characterization protocol

Based on the results of the exploratory analysis (Sections 4.1.1–4.1.4), the following protocol (denoted IP_{LDA}) was established for LDA-based salmon spoilage characterization:

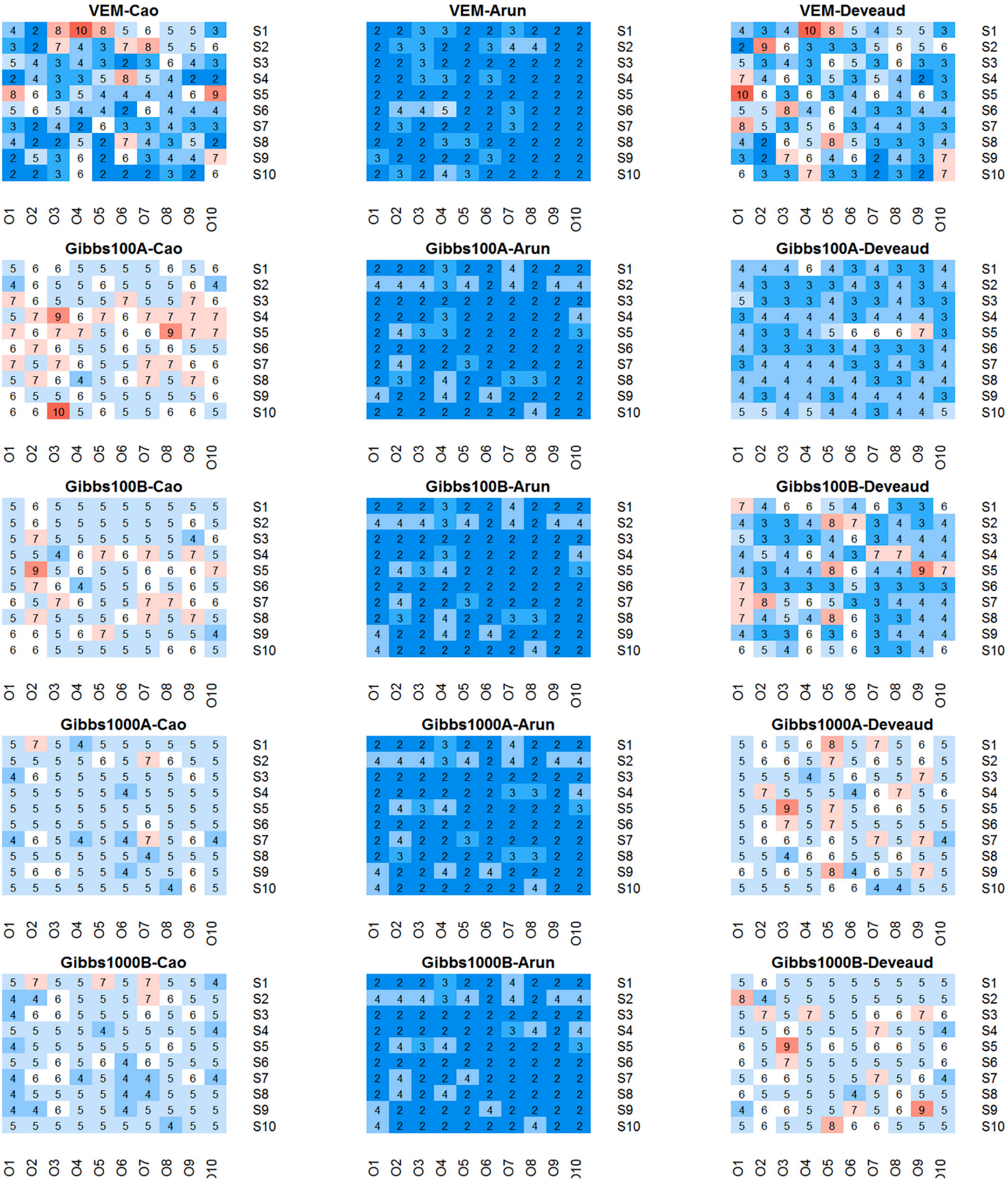
- *Profile number criterion*: for model training, use the lowest k that does not lead to a change in perplexity when compared with the optimal k;
- *Spoilage characterization criterion:* for $SA_{\%}$ calculation, use profiles whose contribution shows a positive correlation with the metadata (here, storage time and $R_{\%}$);
- *Spoilage association criterion*: for identifying potential spoilage indicators, use $SA_{\%} = 0.5$ as the cut-off threshold.

When comparing the performance of the 3-profile IP_{LDA} (denoted IP_{LDA3} in Table 3) with the previously established IP_{PLS}, a high correspondence could be observed at the full protocol level: 5/6 VOCs that fulfilled the three IP_{PLS} selection criteria were also identified by IP_{LDA3}, while 4/5 VOCs having $SA_{\%} = 1$ were also identified by IP_{PLS}. In contrast, a lower correspondence was observed at the methodological level (LDA vs. PLS): out of 15 compounds identified by at least one of the two methods, three compounds (C8, C17, C23) were missed by LDA and three other compounds (C5, C11, C14) by PLS. Finally, a good overall consensus was reached in recognizing irrelevant VOCs: 13/25 compounds were classified as irrelevant by both protocols.

4.2. Selective analysis

The relations between storage time, sensory rejection and profile distribution under conditions B-D are visualized in Fig. 5. Overall, the differences in the evolution of profile distributions showed that the applied atmosphere had a major impact on the progression of spoilage. Under air (B), 60/0/40 (C) and 60/40/0 (D), profiles B3P2, C3P1 and D3P2 could be associated with spoilage, respectively, whereas the other extracted profiles showed little correlation with storage time and/or rejection. In conclusion, only the aforementioned three profiles were considered in $SA_{\%}$ calculations.

All identified potential spoilage indicators are shown in Table 3. Again (see Section 4.1.4), correspondence between the two identification approaches was found to be higher at the protocol level than at the methodological level. Out of 9, 4 and 1 compounds identified by IP_{PLS} under conditions B-D, respectively, 8, 3 and 1 were also identified by IP_{LDA}. Additional compounds identified by IP_{LDA} were 2,3-butanediol, acetone, acetoin, methyl mercaptan and ethyl acetate (C), and ethanol, acetoin, 2,3-butanedione, carbon disulfide, dimethyl sulfide, hydrogen sulfide and ethyl acetate (D). Out of these twelve additional identifications, six were also achieved by PLS.



VEM-Cao

	O1	O2	O3	O4	O5	O6	O7	O8	O9	O10
S1	4	2	8	10	8	5	6	5	5	3
S2	3	2	7	4	3	7	8	5	5	6
S3	5	4	3	4	3	2	3	6	4	3
S4	2	4	3	3	5	8	5	4	2	2
S5	8	6	3	5	4	4	4	4	6	9
S6	5	6	5	4	4	2	6	4	4	4
S7	3	2	4	2	6	3	3	4	3	3
S8	4	2	2	5	2	7	4	3	5	2
S9	2	5	3	6	2	6	3	4	4	7
S10	2	2	3	6	2	2	2	3	2	6

VEM-Arun

	O1	O2	O3	O4	O5	O6	O7	O8	O9	O10
S1	2	2	3	3	2	2	3	2	2	2
S2	2	3	3	2	2	3	4	4	2	2
S3	2	2	3	2	2	2	2	2	2	2
S4	2	2	3	3	2	3	2	2	2	2
S5	2	2	2	2	2	2	2	2	2	2
S6	2	4	4	5	2	2	3	2	2	2
S7	2	3	2	2	2	2	3	2	2	2
S8	2	2	2	3	3	2	2	2	2	2
S9	3	2	2	2	2	3	2	2	2	2
S10	2	3	2	4	3	2	2	2	2	2

VEM-Deveaud

	O1	O2	O3	O4	O5	O6	O7	O8	O9	O10
S1	4	3	4	10	8	5	4	5	5	3
S2	2	9	6	3	3	3	5	6	5	6
S3	5	3	4	3	6	5	3	6	3	3
S4	7	4	6	3	5	3	5	4	2	3
S5	10	6	3	6	3	4	6	4	6	3
S6	5	5	8	4	6	4	3	3	4	4
S7	8	5	3	5	6	3	4	4	3	3
S8	4	2	6	5	8	5	3	3	3	4
S9	3	2	7	6	4	6	2	4	3	7
S10	6	3	3	7	3	3	2	3	2	7

Gibbs100A-Cao

	O1	O2	O3	O4	O5	O6	O7	O8	O9	O10
S1	5	6	6	5	5	5	5	6	5	6
S2	4	6	5	5	6	5	5	5	6	4
S3	7	6	5	5	5	7	5	5	7	6
S4	5	7	9	6	7	6	7	7	7	7
S5	7	6	7	7	5	6	6	9	7	7
S6	6	7	6	5	5	6	5	6	5	5
S7	7	5	7	6	5	5	7	7	6	6
S8	5	7	6	4	5	6	7	5	7	6
S9	6	5	5	6	5	5	5	5	5	6
S10	6	6	10	5	6	5	5	6	6	5

Gibbs100A-Arun

	O1	O2	O3	O4	O5	O6	O7	O8	O9	O10
S1	2	2	2	3	2	2	4	2	2	2
S2	4	4	4	3	4	2	4	2	4	4
S3	2	2	2	2	2	2	2	2	2	2
S4	2	2	2	3	2	2	2	2	2	4
S5	2	4	3	3	2	2	2	2	2	3
S6	2	2	2	2	2	2	2	2	2	2
S7	2	4	2	2	3	2	2	2	2	2
S8	2	3	2	4	2	2	3	3	2	2
S9	4	2	2	4	2	4	2	2	2	2
S10	2	2	2	2	2	2	2	4	2	2

Gibbs100A-Deveaud

	O1	O2	O3	O4	O5	O6	O7	O8	O9	O10
S1	4	4	4	6	4	3	4	3	3	4
S2	4	3	3	3	3	4	3	4	3	4
S3	5	3	3	3	4	3	3	4	3	3
S4	3	4	4	4	4	3	4	4	4	3
S5	4	3	3	4	5	6	6	6	7	3
S6	4	3	3	3	3	4	3	3	3	4
S7	3	4	4	4	4	3	3	4	3	4
S8	4	4	4	4	4	4	3	3	4	4
S9	4	3	4	4	3	4	4	4	4	3
S10	5	5	4	5	4	4	3	4	4	5

Gibbs100B-Cao

	O1	O2	O3	O4	O5	O6	O7	O8	O9	O10
S1	5	6	5	5	5	5	5	5	5	5
S2	5	6	5	5	5	5	5	5	6	5
S3	5	7	5	5	5	5	5	5	4	6
S4	5	5	4	6	7	6	7	5	7	5
S5	5	9	5	6	5	5	6	6	6	7
S6	5	7	6	4	5	5	6	5	6	5
S7	6	5	7	6	5	5	7	7	6	6
S8	5	7	5	5	5	6	7	5	7	5
S9	6	6	5	6	7	5	5	5	5	4
S10	6	6	5	5	5	5	5	6	6	5

Gibbs100B-Arun

	O1	O2	O3	O4	O5	O6	O7	O8	O9	O10
S1	2	2	2	3	2	2	4	2	2	2
S2	4	4	4	3	4	2	4	2	4	4
S3	2	2	2	2	2	2	2	2	2	2
S4	2	2	2	3	2	2	2	2	2	4
S5	2	4	3	4	2	2	2	2	2	3
S6	2	2	2	2	2	2	2	2	2	2
S7	2	4	2	2	3	2	2	2	2	2
S8	2	3	2	4	2	2	3	3	2	2
S9	4	2	2	4	2	4	2	2	2	2
S10	4	2	2	2	2	2	2	4	2	2

Gibbs100B-Deveaud

	O1	O2	O3	O4	O5	O6	O7	O8	O9	O10
S1	7	4	6	6	5	4	6	3	3	6
S2	4	3	3	4	8	7	3	4	3	4
S3	5	3	3	3	4	6	3	4	4	4
S4	4	5	4	6	4	3	7	7	4	4
S5	4	3	4	4	8	6	4	4	9	7
S6	7	3	3	3	3	5	3	3	3	3
S7	7	8	5	6	5	3	3	4	4	4
S8	7	4	5	4	8	6	3	3	4	4
S9	4	3	3	6	3	6	3	4	4	4
S10	6	5	4	6	5	6	3	3	4	6

Gibbs1000A-Cao

	O1	O2	O3	O4	O5	O6	O7	O8	O9	O10
S1	5	7	5	4	5	5	5	5	5	5
S2	5	5	5	5	6	5	7	6	5	5
S3	4	6	5	5	5	5	5	5	6	5
S4	5	5	5	5	5	4	5	5	5	5
S5	5	5	5	5	5	5	5	5	5	5
S6	5	5	5	5	5	5	6	5	5	5
S7	4	6	5	4	5	4	7	5	6	4
S8	5	5	5	5	5	5	4	5	5	5
S9	5	6	6	5	5	4	5	5	6	5
S10	5	5	5	5	5	5	5	4	6	5

Gibbs1000A-Arun

	O1	O2	O3	O4	O5	O6	O7	O8	O9	O10
S1	2	2	2	3	2	2	4	2	2	2
S2	4	4	4	3	4	2	4	2	4	4
S3	2	2	2	2	2	2	2	2	2	2
S4	2	2	2	2	2	2	3	3	2	4
S5	2	4	3	4	2	2	2	2	2	3
S6	2	2	2	2	2	2	2	2	2	2
S7	2	4	2	2	3	2	2	2	2	2
S8	2	3	2	2	2	2	3	3	2	2
S9	4	2	2	4	2	4	2	2	2	2
S10	4	2	2	2	2	2	2	4	2	2

Gibbs1000A-Deveaud

	O1	O2	O3	O4	O5	O6	O7	O8	O9	O10
S1	5	6	5	6	8	5	7	5	6	5
S2	5	6	6	5	7	5	6	5	6	5
S3	5	5	5	4	5	6	5	5	7	5
S4	5	7	5	5	5	4	6	7	5	6
S5	5	5	9	5	7	5	6	6	5	5
S6	5	6	7	5	7	5	5	5	5	5
S7	5	6	6	5	6	5	7	5	7	4
S8	5	5	4	6	6	5	5	6	5	5
S9	6	5	6	5	8	4	6	5	7	5
S10	5	5	5	5	6	6	4	4	5	5

Gibbs1000B-Cao

	O1	O2	O3	O4	O5	O6	O7	O8	O9	O10
S1	5	7	5	5	7	5	7	5	5	4
S2	4	4	6	5	5	5	7	6	5	5
S3	4	6	6	5	5	5	6	5	6	5
S4	5	5	5	5	4	5	5	5	5	4
S5	4	5	5	5	5	5	5	6	5	5
S6	5	5	6	5	6	4	6	5	5	5
S7	4	6	6	4	5	4	4	5	6	4
S8	4	5	5	5	5	4	4	5	5	5
S9	4	4	6	5	5	4	5	5	5	5
S10	5	5	5	5	5	5	5	4	5	5

Gibbs1000B-Arun

	O1	O2	O3	O4	O5	O6	O7	O8	O9	O10
S1	2	2	2	3	2	2	4	2	2	2
S2	4	4	4	3	4	2	4	2	4	4
S3	2	2	2	2	2	2	2	2	2	2
S4	2	2	2	2	2	2	3	4	2	4
S5	2	4	3	4	2	2	2	2	2	3
S6	2	2	2	2	2	2	2	2	2	2
S7	2	4	2	2	4	2	2	2	2	2
S8	2	4	2	4	2	2	2	2	2	2
S9	4	2	2	2	2	4	2	2	2	2
S10	4	2	2	2	2	2	2	4	2	2

Gibbs1000B-Deveaud

	O1	O2	O3	O4	O5	O6	O7	O8	O9	O10
S1	5	6	5	5	5	5	5	5	5	5
S2	8	4	5	5	5	5	5	5	5	5
S3	5	7	5	7	5	5	6	6	7	6
S4	5	5	6	5	5	5	7	5	5	4
S5	6	5	9	5	6	5	6	6	5	6
S6	6	5	7	5	5	5	5	5	5	6
S7	5	6	6	5	5	5	7	5	6	4
S8	6	5	5	5	5	4	5	6	5	5
S9	4	6	6	5	5	7	5	6	9	5
S10	5	6	5	5	8	6	6	5	5	5

Fig. 2. The optimal number of profiles (k), extracted from the volatilome of Atlantic salmon fillet portions stored under 100% N_2 (condition A) at 4 °C. For methodological specifications concerning the applied metrics (Cao, Arun, Deveaud), inference estimation methods (VEM, Gibbs100A, Gibbs100B, Gibbs1000A and Gibbs1000B), seeds (s1-s10) and input column orders (O1-O10), see Section 3.2.1.

Table 2
The perplexities of the Latent Dirichlet Allocation (LDA) models of salmon spoilage under condition A (100% N_2), resulting from leave-one-out cross validation (listed according to holdout sample ID).

ID	Time (d)	Number of profiles								
		2	3	4	5	6	7	8	9	10
1	1	15.14	7.96	7.85	7.83	7.60	7.72	7.70	7.76	7.66
2	1	29.18	10.41	10.35	10.33	10.34	10.34	10.31	10.30	10.35
3	1	18.71	15.13	8.04	7.97	7.90	7.82	7.84	7.87	7.86
4	3	38.58	11.81	11.10	10.95	10.99	11.04	10.98	11.08	11.13
5	3	52.30	12.81	12.67	12.78	12.81	12.86	12.89	12.87	12.99
6	5	2.22	2.19	2.16	2.16	2.16	2.16	2.15	2.16	2.15
7	5	2.30	2.28	2.28	2.28	2.27	2.28	2.28	2.27	2.27
8	5	3.42	3.34	3.18	3.17	3.16	3.17	3.16	3.16	3.17
9	7	1.92	1.92	1.92	1.92	1.92	1.92	1.91	1.91	1.91
10	7	1.81	1.81	1.80	1.80	1.80	1.80	1.80	1.80	1.80
11	9	1.80	1.81	1.78	1.78	1.78	1.78	1.78	1.78	1.78
12	9	1.83	1.83	1.83	1.83	1.83	1.83	1.83	1.83	1.83
13	9	2.19	2.17	2.16	2.16	2.16	2.16	2.16	2.16	2.16
14	11	2.02	2.00	1.99	1.98	1.98	1.98	1.98	1.98	1.98
15	11	3.63	3.49	3.60	3.52	3.49	3.46	3.45	3.48	3.37
16	13	2.34	2.33	2.29	2.27	2.26	2.26	2.26	2.26	2.26

5. Discussion

5.1. Model optimization

In machine learning and statistical analysis, tuning refers to a process where different parameters are tested in order to optimize model performance. Even though it is generally well known that tuning may greatly affect the LDA output – for example, Agrawal et al. (2018) concluded that neither reusing the tunings of a preceding study nor relying on "off-the-shelf" settings can be recommended – relatively few efforts of evaluating and/or controlling the impact of LDA tuning on model output and performance have been published so far. For this reason, special emphasis was given in the present study on a systematic selection of appropriate tunings. The key points of this decision-making process are elaborated in the following paragraphs.

An inference algorithm is needed for approximating the inference of the posterior distribution. The two options considered in the present study – VEM and Gibbs sampling – represent two widely popular and yet fundamentally different approaches. Unlike VEM, Gibbs sampling does not converge to a point estimate but generates random samples from a complex distribution, meaning that the true distribution will be eventually reached (Binkley et al., 2014). The fact that VEM resulted in a high variation between individual models (Fig. 2) was thus not unexpected, as it was likely due to converging towards different local maxima. It must be emphasized though that the choice of Gibbs sampling over VEM does not guarantee finding the true distribution per se, as ensuring the representability of the obtained results in the former case requires additional attention on the *Gibbs sampling parameters* (denoted here as *iter*, *burnin* and *thin*). Briefly, a sufficiently high number of burned iterations is needed for ensuring that the sampler converges to the correct distribution, whereas thinning has traditionally been considered advantageous in addressing the risks associated with autocorrelation (Binkley et al., 2014). However, with regard to the remarks of Link and Eaton (Link and Eaton, 2012) on the thinning of Markov Chain Monte Carlo (MCMC) chains, a 10-fold increase in the number of iterations (100 vs. 1000) was used instead of thinning for reducing the risk of autocorrelation in the present study. Given the achieved level of stabilization in non-burned chains and the additional beneficial impact of burning (Section 4.1.1), the method Gibbs1000B (*iter* = 1000, *burnin* = 500, *thin* = 1) was considered appropriate for further modelling activities.

Irrespective of the used inference algorithm, factors such as *Dirichlet hyperparameters* and *the order of input (WDM) columns* should be taken into account. The hyperparameters α and β represent the Dirichlet priors on the output distributions (Binkley et al., 2014); a high value of α (β, respectively) indicates that a given document (topic, resp.) is likely to consist of a broad range of topics (words, resp.), whereas a low value suggests that only a few topics (words, resp.) are involved. The relevance of these parameters has recently been highlighted by Park et al. (2019), emphasizing that the underlying assumption of unimodality may lead to biased parameter estimation if the corpus consists of clusters with different topic distributions. However, since 1) all extracted profiles can initially be assumed to comprise all VOCs and to be present in each sample, but 2) no prior assumptions should be made on the exact composition and/or distribution of these profiles, the default settings of the packaging **topicmodels** were considered appropriate in this study. On the other hand, since 1) the order of VOCs within the WDM can and must be considered fully exchangeable, but 2) a high level of stabilization was achieved with the method Gibbs1000B (Section 4.1.1), examining a single random seed-order pair was considered sufficient.

In this study, the impact of model tuning on the *number of extracted profiles* was assessed by means of three popular metrics (Section 3.2.1). Since no overall consensus was reached (Section 4.1.1), cross-validation was performed. The fact that the highest perplexity coincided with the beginning of storage time (days 1–3) was attributed to the characteristic deterioration patterns of salmon under the tested storage condition (100% N_2). In the absence of both CO_2 and O_2, considerable accumulation of ethanol and sulfuric compounds was observed over storage time (Kuuliala et al., 2019); for this reason, the relatively small proportion of early-day samples (days 1–3) in a given training set resulted in a better fit for the late-day samples (days 5–13). However, apart from the ill-performing two-profile models, little difference in perplexity was observed between models with different values of k. For this reason, three (the smallest possible k according to perplexity), five (Md(k) according to the Cao and Deveaud metrics) and nine (an example of a high k) profiles were chosen for further exploratory activities.

Finally, *computational power* may pose additional challenges during exploratory analysis. As the order of extracted topics is exchangeable even between consecutive iterations (Binkley et al., 2014), the optimization of model parameters for a specific experimental setting requires comparison between multiple corresponding but independent models. While learning a single LDA model did not take considerably longer than PLS, the total time needed for generating 100 models (10 orders × 10 seeds) with a given method ranged from 23 min (VEM) to over 8.5 h (Gibbs1000B). This was anyhow considered acceptable at the exploratory stage, not only because of the benefits over PLS (see Sections 2 and 5.2) but also because of the positive impact on the analytical workflow. In other words, sufficient emphasis on model tuning at the exploratory stage considerably reduced the need for computational effort during the selective stage.

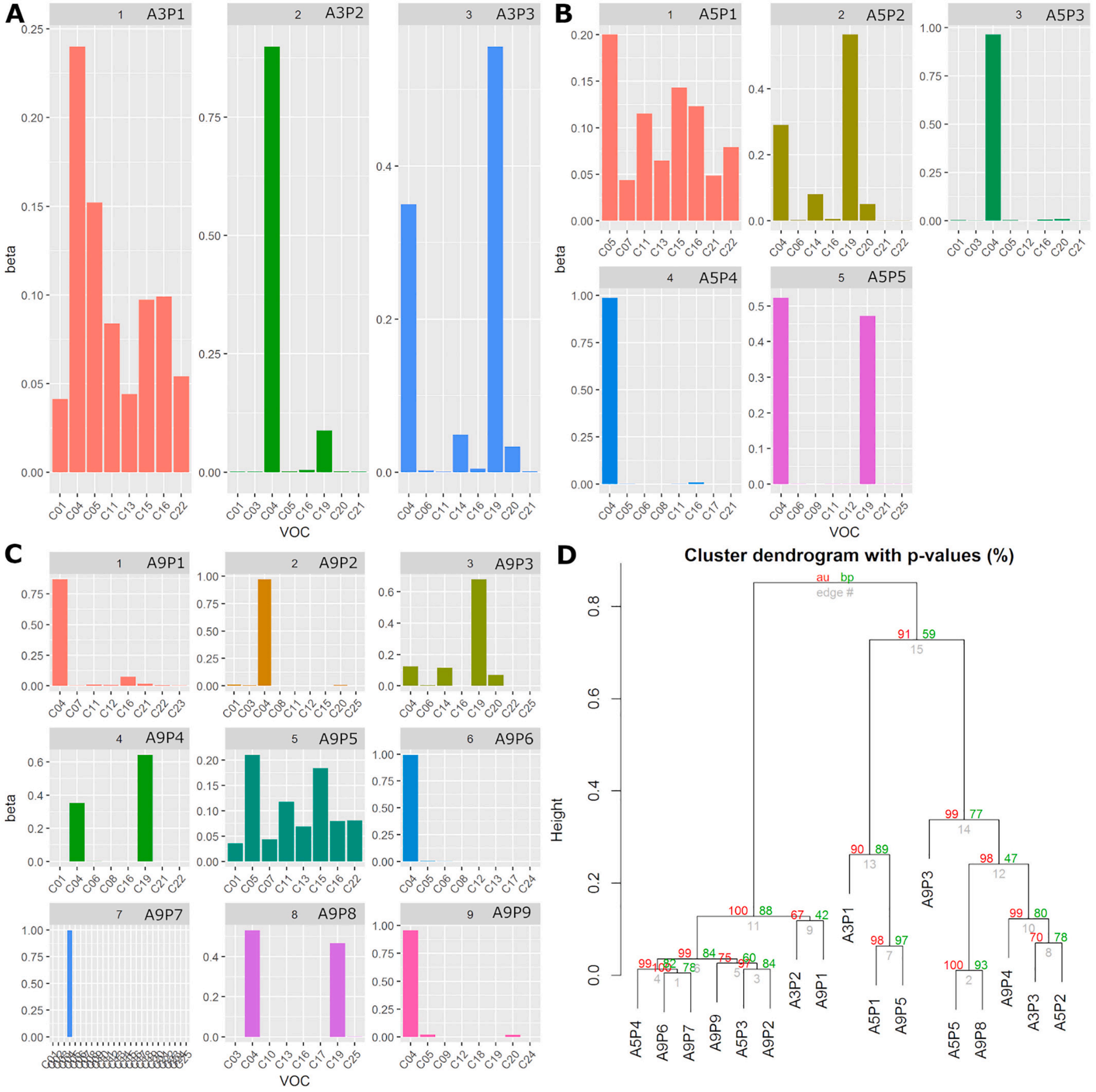


Fig. 3. The distribution of the top-8 volatiles in the profiles extracted from the volatilome of Atlantic salmon fillet portions stored under 100% N_2 (condition A) at 4 °C; A) $k = 3$, B) $k = 5$, C) $k = 9$, and D) hierarchical clustering of all profiles.

5.2. Spoilage characterization

Despite the increasing popularity of topic modelling in scientific discovery, information about its biological applications is still rather scarce in the current literature. Since the publication of the review of Liu et al. (2006), a limited number of studies have been published on topics such as genetic/protein functionality (Backenroth et al., 2018; Liu et al., 2017; Liu et al., 2018), metabolic sub-structures (van der Hooft et al., 2016) and mutation signatures (Matsutani et al., 2019). The present study thus extends the current state-of-the-art by introducing LDA as the basis of a systematic spoilage characterization protocol. The central focus points that should be considered when applying LDA as an exploratory and/or selective method within this specific context are discussed in the following paragraphs; for further information about the identified compounds and their role in seafood spoilage, please see the preceding study of Kuuliala et al. (2019).

Firstly, the method of *profile interpretation* greatly determines the performance and representability of LDA. This highlights the importance of metadata in spoilage analysis, particularly if no prior knowledge about the expected deterioration processes and/or potential spoilage indicators exists. Hence, storage time and sensory rejection were used in the present study for confirming the tentative interpretations arising from the VOC distributions per se. For example, the correlations between increasing rejection, decreasing contribution of cluster 1 and subsequent increasing contributions of clusters 2–3 under 100% N_2 (Section 4.1.3) indicated that the former cluster could be

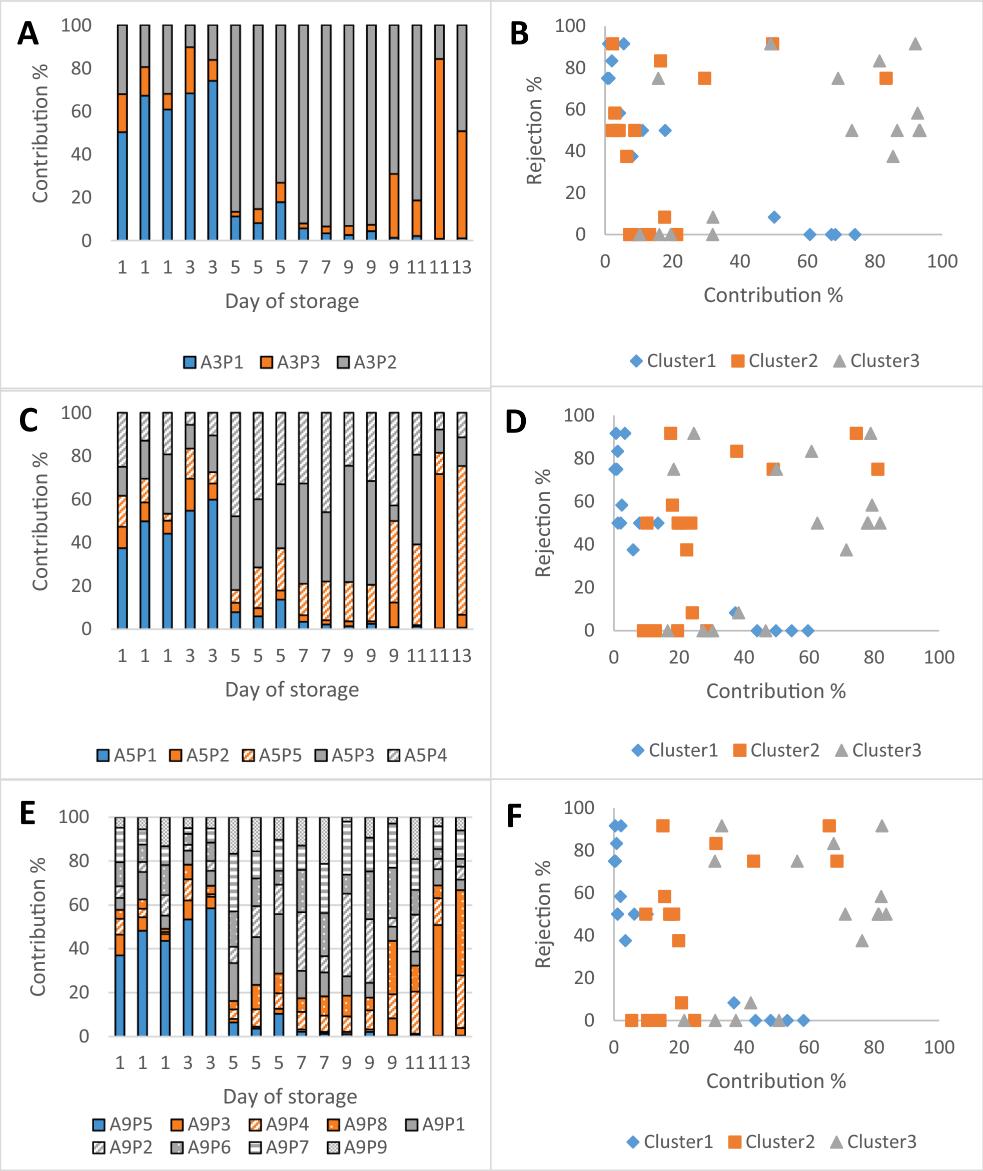


Fig. 4. The relations between storage time (from left to right: sample ID 1–16), sensory rejection ($R_{\%}$), and profile/cluster distribution in Atlantic salmon fillet portions stored under 100% N_2 (condition A) at 4 °C; A-B) k = 3, C-D) k = 5, E-F) k = 9.

associated with freshness and the latter two with spoilage. Furthermore, the emergence of multiple spoilage-associated profiles under condition A (Fig. 4) and those showing stable or fluctuating patterns under conditions B-D (Fig. 5) suggest that the progression of spoilage cannot necessarily be comprehensively modelled with a single profile, at least in case of small datasets and/or in the absence of extremely fresh/spoiled samples.

The number of extracted profiles has both mathematical and biological relevance. Even though all extracted profiles and their relations provide information about salmon quality and can thus be considered relevant for exploratory purposes, a systematic selection criterion was anyhow needed for selective analysis. In line with the previous PLS-based protocol (Kuuliala et al., 2019), choosing the lowest k (condition A: 3) that does not lead to an increase in perplexity when compared with the optimal k (condition A: 5) was considered appropriate. Furthermore, additional perplexity analyses performed for conditions B-D (results not shown) indicated that different holdout sample ID and/or k had little impact on perplexity; under elevated CO_2 and/or O_2 levels, the final concentrations were considerably lower and the relations between VOCs more stable when compared to storage under 100% N_2 (Kuuliala et al., 2019), meaning that the differences between all samples (and thus between the extracted profiles) were less pronounced when compared to those under condition A (see Section 5.1). Consequently, k = 3 was consistently used for selective purposes under all tested conditions.

When aiming at using LDA for selective analysis, relevant *definitions* should be established in the first place. In the case of spoilage analysis, this essentially means determining what kind of VOCs can be considered spoilage indicators. Firstly, it should be noted that all VOCs present at a given time point do not necessarily contribute to the perceived off-odors, as this requires exceeding certain concentration thresholds. However, it is equally important to note that these thresholds are both compound-specific and context-dependent. For example, while the human olfactory threshold (OT) of a given VOC indicates its minimum perceivable concentration (Devos et al., 1990), the previously reported OTs have usually been defined for pure compounds, whereas the seafood volatilome consists of multiple compounds which may interact or interfere with each other. Consequently, exceeding the OT does not guarantee that a VOC can be perceived as a part of a complex volatilome or that it

Table 3
A comparison of potential spoilage indicators in Atlantic salmon (Salmo salar) stored under gaseous atmospheres (% $CO_2/O_2/N_2$) A (0/0/100), B (air), C (60/0/40) and D (60/40/0), identified by protocols based on Latent Dirichlet Allocation (IP_{LDA}[a]) or partial least squares regression (IP_{PLS}[b]).

VOC	A				B		C		D	
	IP_{LDA3}	IP_{LDA5}	IP_{LDA9}	IP_{PLS}	IP_{LDA3}	IP_{PLS}	IP_{LDA3}	IP_{PLS}	IP_{LDA3}	IP_{PLS}
Acetic acid (C1)	0.5	0.25	0.375	x	0	x	0	x	0	x
3-Methylbutanoic acid (C2)	0	0	0		0		0		0	
2,3-Butanediol (C3)	0.5	0.25	0.375	**X**	0	**X**	1		0	
Ethanol (C4)	1	1	1	**X**	1	**X**	1	**X**	1	
3-Methyl-1-butanol (C5)	0.5	0.5	0.25		1	**X**	0		0	
Isobutyl alcohol (C6)	0.5	0.75	0.5	x	1	**X**	0		0	x
3-Methylbutanal (C7)	0	0	0.125		1	**X**	1	**X**	1	**X**
Ethyl benzene (C8)	0	0.25	0.375	x	0	x	0		0	
Propyl benzene (C9)	0	0.25	0.125		0		0		0	
Styrene (C10)	0	0	0.125		0		0		0	
Acetone (C11)	0.5	0.5	0.375		1	**X**	1		0	x
Acetoin (C12)	0	0.5	0.5	**X**	1	**X**	1	x	1	x
2,3-Butanedione (C13)	0	0	0.25		0		0		1	x
Butanone (C14)	0.5	0.25	0.25		0		0		0	
Carbon disulfide (C15)	0	0	0		0		0		1	
Dimethyl sulfide (C16)	1	0.75	0.625	**X**	1	**X**	0	**X**	1	
Dimethyl disulfide (C17)	0	0.25	0.25	x	0		0		0	
Dimethyl trisulfide (C18)	0	0	0.125		0		0		0	
Hydrogen sulfide (C19)	1	0.5	0.625	**X**	0		1	**X**	1	
Methyl mercaptan (C20)	1	0.5	0.375	**X**	0	x	1	x	0	x
Ethyl acetate (C21)	1	1	0.25	x	1	**X**	1	x	1	x
Ammonia (C22)	0	0.25	0.375		0		0		0	
Dimethylamine (C23)	0	0	0.125	x	0	x	0	x	0	
Piperidine(C24)	0	0	0.25		0		0		0	
Trimethylamine (C25)	0	0.25	0.375		0		0		0	

[a] IP_{LDAk}: identification based on relative spoilage contributions ($SC_{\%}$; 0–1) in a k-profile LDA model (here, $k = 3, 5, 9$).

[b] IP_{PLS}: identification based on the fulfilment of Cr-3 (x) or all three criteria (**X**) of Kuuliala et al. (2019).

contributes to offensive off-odors: the thresholds associated with these two aspects are in fact often not well known before the commence of the intended study. Anyhow, whether this matters depends on the scientific and/or practical context. As previously highlighted (Ioannidis et al., 2018; Kuuliala et al., 2019), a VOC that shows a high positive correlation with microbiological and/or sensory deterioration may have great value in quality monitoring applications even if its individual olfactory contribution remains low or unknown. For these reasons, the concept "potential spoilage indicator" is used in the present study to refer to all VOCs that show promising potential in monitoring quality decay, irrespective of their individual olfactory contribution.

Along with model optimization (Section 5.1), the methods of *data pre-processing* have a key role in ensuring that the applied methodology is in line with the defined aims. The relevance of this remark can be seen when comparing the performance of PLS and LDA (Table 3). In the former case, the data had been standardized in order to disregard the differences in concentration magnitudes between individual VOCs; furthermore, additional data quality criteria were implemented at the identification stage to dismiss VOCs whose quantification accuracy was considered inadequate (for more information, see Kuuliala et al., 2019). In contrast, all LDA models were learned with non-standardized data and used without additional data quality screening. The high correspondence between the final outcomes of the two protocols – despite the aforementioned methodological differences – was thus attributed to the characteristics of the salmon datasets. As elaborated in the previous study (Kuuliala et al., 2019), the data quality criteria that were used for screening VOCs primarily targeted those compounds that were present in low concentrations (< 100 ppb) throughout storage time; due to the lack of standardization, this kind of compounds (for example, all quantified amines) typically had little impact on the extracted LDA profiles. However, the lack of data quality screening also increased the number of LDA-based identifications of low-range VOCs when compared to PLS, particularly under condition D where the differences in concentration magnitudes between individual VOCs were at the lowest. Overall, these results demonstrate that while LDA is less sensitive to problems arising from the presence of low-range VOCs than PLS, an initial data quality analysis can be generally considered advisable.

For finalizing the selective analysis, representative *cut-off thresholds* are needed. In case of spoilage analysis, it is of primary importance to note that the sole presence/absence of a VOC does not automatically signify a certain quality status: instead, VOC-based spoilage characterization requires considering both the evolution and relations of all quantified compounds throughout storage time. For this reason, the concept of relative spoilage association was developed in this study to select those VOCs which had the highest (top-8) relative abundance within spoilage-associated profiles. In line with the definition of a potential spoilage indicator (see above), this concept was developed for quality monitoring purposes and does thus not directly signify individual olfactory contribution. For example, only a single spoilage-associated profile per condition could be identified under conditions B-D (Section 4.2), meaning that a given VOC could only receive an $SA_{\%}$ value 0 (irrelevant) or 1 (relevant). In the case of condition A, multiple spoilage-associated profiles were identified, meaning that a well-established numerical cut-off limit was needed. The commonly observed increase/decrease in $SA_{\%}$ as a function of k (Table 3) was attributed to sub-profile partitioning (Section 4.1.3), which increased the overall diversity of the top-8 VOCs and thus reduced the number of compounds with extreme $SA_{\%}$ values (0 or 1). In this study, setting the limit at $SA_{\%} = 0.5$ was experimentally found to respond to the aforementioned needs in an optimal manner as well as to result in a high correspondence between the two identification protocols (IP_{LDA} and IP_{PLS}); however, it is important to note that the selection of the cut-off limit should always be done on a case-by-case basis.

Finally, it is advisable to evaluate *the prospects* of a newly developed methodology in a broader context. In this case, attention was thus given to the applicability of LDA in the volatilome-based spoilage characterization of other food products. Generally, the developed methodology is expected to be widely applicable for examining the evolution and composition of complex volatilomes, suggesting that highly perishable products packed under gaseous atmospheres (such as vegetables, meat and seafood) whose quality decay is manifested by the accumulation of multiple VOCs (resulting in unacceptable off-odors) would be the most

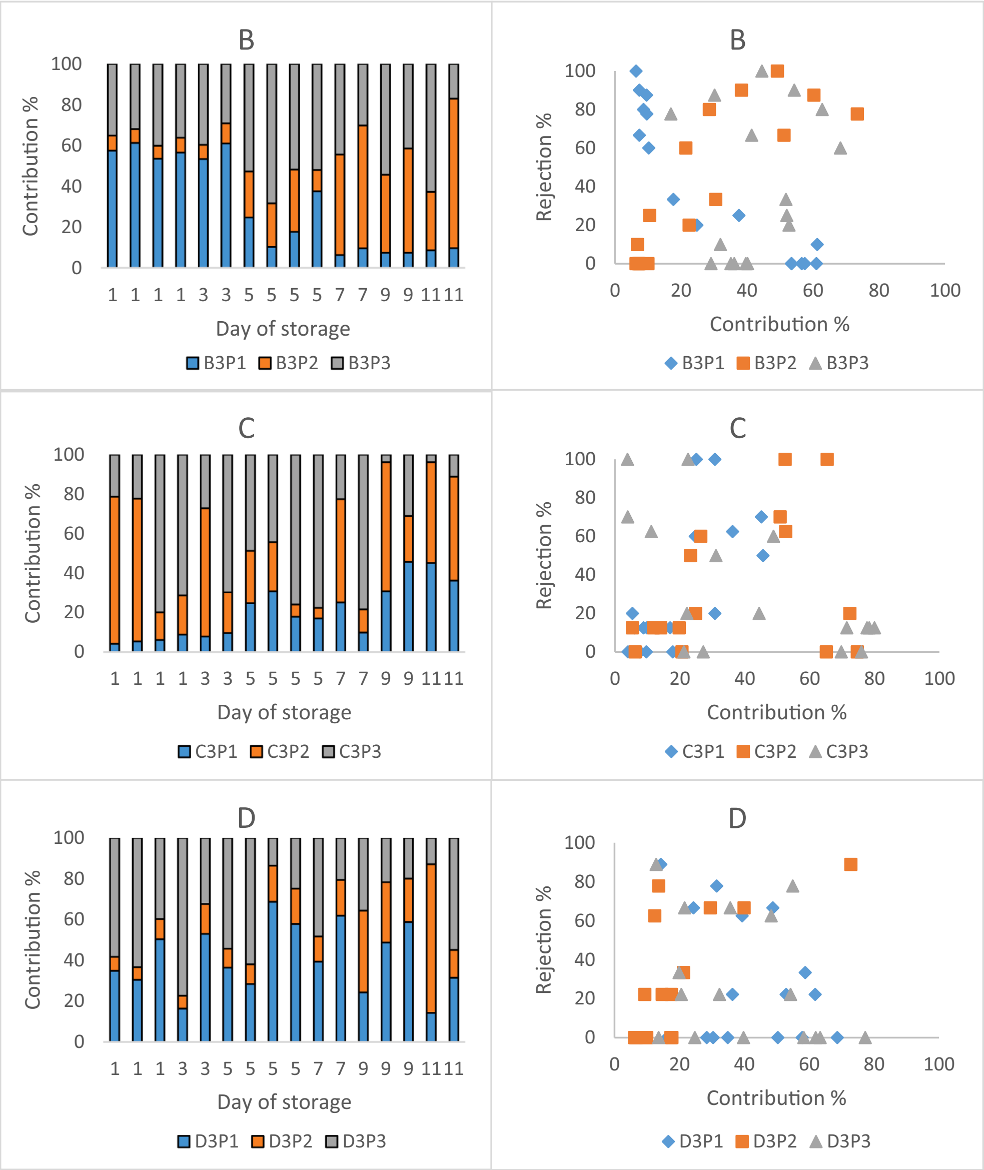


Fig. 5. The relations between storage time (from left to right: sample ID 1–16), sensory rejection ($R_{\%}$) and profile distribution (k = 3) in Atlantic salmon fillet portions stored under different gaseous conditions (% $CO_2/O_2/N_2$) at 4 °C; air (B), 60/0/40 (C) and 60/40/0 (D).

promising target group. While model tunings and cut-off limits should be experimentally determined whenever introducing a new product, the basic development process described in the present study could be used as a theoretical basis for exploring corresponding research questions in new experimental settings. In the future, emphasis should thus be given not only on testing and validating the method in these settings, but also on product-specific planning that precedes actual modelling activities. Preferably, the availability and quality of the input data should be considered already before setting up a storage experiment: since each food product has its characteristic shelf-life, the experimental setup should allow regular and representative data collection throughout storage time. Before selective analysis, the relation between the VOCs and the dependent variable should receive particular attention, as it is not always linear: for example, reaching 100% rejection does not mean that the volatilome will also be stable from there on. In general, it should be kept in mind that the prevailing assumptions regarding the complex quality deterioration mechanisms may affect the entire spoilage characterization process: the fact that the training of LDA models does not inherently involve metadata could thus be considered advantageous, especially when considering new product types. For further insights, a comparison between unsupervised and supervised LDA would be worth examination.

6. Conclusions

The outcomes of the present study show that LDA can be successfully applied for extracting underlying information about the quality status of Atlantic salmon under different storage conditions, suggesting that a respective approach could be well adapted for other food products and/or quality deterioration patterns. As a selective tool, LDA was found to identify potential volatile spoilage indicators with equal specificity and lower stringency when compared to PLS. In particular, the flexibility and high interpretability of LDA were considered highly advantageous in this problem setting; not only because of their beneficial impact on the experimental workflow, but also because of the achieved insights into the development of systematic spoilage characterization processes. Overall, the results support the state-of-the-art of data-driven food quality characterization, an emerging field with great prospects.

Declaration of competing interest

None.

Acknowledgments

This research article was realized in the framework of the TERAFOOD project supported by the European Regional Development Fund and the province Oost-Vlaanderen, the “Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen” programme supported by the Flemish Government, and the CheckPack project (IWT-SBO-130036) supported by Flanders Innovation & Entrepreneurship (VLAIO, formerly known as the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT)) and by a diverse group of industrial stakeholders within the packaging industry. L. Kuuliala acknowledges the support of the Research Foundation Flanders for a post-doctoral fellow (1222020N). R. Pérez-Fernández acknowledges the support of the Research Foundation Flanders for a post-doctoral fellow (FWO17/PDO/160) and the Spanish MINECO (TIN2017-87600-P). M. Tang is supported by the China Scholarship Council (CSC).

References

Agrawal, A., Fu, W., Menzies, T., 2018. What is wrong with topic modeling? And how to fix it using search-based software engineering. Inf. Softw. Technol. 98, 74–88.

Arun, R., Suresh, V., Veni Madhavan, C.E., Narasimha Murthy, M.N., 2010. On finding the natural number of topics with latent dirichlet allocation: some observations. In: Zaki, M.J., Yu, J.X., Ravindran, B., Pudi, V. (Eds.), Advances in Knowledge Discovery and Data Mining. PAKDD 2010. Lecture Notes in Computer Science, vol. 6118. Springer, Berlin, Heidelberg.

Backenroth, D., He, Z., Kiryluk, K., Boeva, V., Pethukova, L., Khurana, E., Christiano, A., Buxbaum, J., Ionita-Laza, I., 2018. FUN-LDA: a Latent Dirichlet Allocation model for predicting tissue-specific functional effects of noncoding variation: methods and applications. Am. J. Hum. Genet. 102, 920–942.

Bastani, K., Namavari, H., Shaffer, J., 2019. Latent Dirichlet allocation (LDA) for topic modeling of the CFPB consumer complaints. Expert Syst. Appl. 127, 256–271.

Bermejo-Prada, A., Vega, E., Pérez-Mateos, M., Otero, L., 2015. Effect of hyperbaric storage at room temperature on the volatile profile of strawberry juice. LWT - Food Sci. Technol. 62, 906–914.

Binkley, D., Heinz, D., Lawrie, D., Overfelt, J., 2014. Understanding LDA in source code analysis. In: In: Proceedings of the 22nd International Conference on Program Comprehension, ACM, Hyderabad, India, pp. 26–36.

Blei, D.M., 2012. Probabilistic topic models. Commun. ACM 55, 77–84.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent Dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022.

Böhme, K., Calo-Mata, P., Barros-Velázquez, J., Ortea, I., 2019. Recent applications of omics-based technologies to main topics in food authentication. TrAC Trend. Anal. Chem. 110, 221–232.

Cao, J., Xia, T., Li, J., Zhang, Y., Tang, S., 2009. A density-based method for adaptive LDA model selection. Neurocomputing 72, 1775–1781.

Castro-Puyana, M., Pérez-Míguez, R., Montero, L., Herrero, M., 2017. Application of mass spectrometry-based metabolomics approaches for food safety, quality and traceability. TrAC Trend. Anal. Chem. 93, 102–118.

Chen, X., Hu, X., Shen, X., Rosen, G., 2010. Probabilistic topic modeling for genomic data interpretation. In: 2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) Hong Kong, pp. 149–152.

Curiskis, S.A., Drake, B., Osborn, T.R., Kennedy, P.J., 2020. An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit. Inform. Process. Manag. 57, 102034.

Deveaud, R., Sanjuan, É., Bellot, P., 2014. Accurate and effective latent concept modeling for ad hoc information retrieval. Document Numérique, Lavoisier 2014, 61–84.

Devos, M., Patte, F., Rouault, J., Laffort, P., Van Gemert, L.J. (Eds.), 1990. Standardized Human Olfactory Thresholds. Oxford University Press, New York, US.

Dong, D., Jiao, L., Li, C., Zhao, C., 2019. Rapid and real-time analysis of volatile compounds released from food using infrared and laser spectroscopy. TrAC Trend. Anal. Chem. 110, 410–416.

Feinerer, I., Hornik, K., 2017. tm: Text Mining Package. R package version 0.7-3. URL. https://CRAN.R-project.org/package=tm.

Feinerer, I., Hornik, K., Meyer, D., 2008. Text mining infrastructure in R. J. Stat. Softw. 25, 1–54.

Fu, Y., Yan, M., Zhang, X., Xu, L., Yang, D., Kymer, J.D., 2015. Automated classification of software change messages by semi-supervised Latent Dirichlet Allocation. Inf. Softw. Technol. 57, 369–377.

Ghasemi-Varnamkhasti, M., Apetrei, C., Lozano, J., Anyogu, A., 2018. Potential use of electronic noses, electronic tongues and biosensors as multisensor systems for spoilage examination in foods. Trends Food Sci. Technol. 80, 71–92.

Griffiths, T.L., Steyvers, M., 2004. Finding scientific topics. Proc. Natl. Acad. Sci. 101, 5228–5235.

Grün, B., Hornik, K., 2011. topicmodels: an R package for fitting topic models. J. Stat. Softw. 40, 1–30.

Hu, N., Zhang, T., Gao, B., Bose, I., 2019. What do hotel customers complain about? Text analysis using structural topic model. Tour. Manag. 72, 417–426.

Ioannidis, A., Kerckhof, F., Riahi Drif, Y., Vanderroost, M., Boon, N., Ragaert, P., De Meulenaer, B., Devlieghere, F., 2018. Characterization of spoilage markers in modified atmosphere packaged iceberg lettuce. Int. J. Food Microbiol. 279, 1–13.

Klampfl, C.W., 2018. Ambient mass spectrometry in foodomics studies. Curr. Opin. Food Sci. 22, 137–144.

Kullback, S., Leibler, R.A., 1951. On information and sufficiency. Ann. Math. Stat. 22, 79–86.

Kuuliala, L., Abatih, E., Ioannidis, A.-G., Vanderroost, M., De Meulenaer, B., Ragaert, P., Devlieghere, F., 2018. Multivariate statistical analysis for the identification of potential seafood spoilage indicators. Food Control 84, 49–60.

Kuuliala, L., Sader, M., Solimeo, A., Pérez-Fernández, R., Vanderroost, M., De Baets, B., De Meulenaer, B., Ragaert, P., Devlieghere, F., 2019. Spoilage evaluation of raw Atlantic salmon (*Salmo salar*) stored under modified atmospheres by multivariate statistics and augmented ordinal regression. Int. J. Food Microbiol. 303, 46–57.

Li, X., Ma, Z., Peng, P., Guo, X., Huang, F., Wang, X., Guo, J., 2018. Supervised latent Dirichlet allocation with a mixture of sparse softmax. Neurocomputing 312, 324–335.

Link, W.A., Eaton, M.J., 2012. On thinning of chains in MCMC. Methods Ecol. Evol. 3, 112–115.

Liu, L., Tang, L., Dong, W., Yao, S., Zhou, W., 2016. An overview of topic modeling and its current applications in bioinformatics. SpringerPlus 5, 1608.

Liu, L., Tang, L., He, L., Yao, S., Zhou, W., 2017. Predicting protein function via multi-label supervised topic model on gene ontology. Biotechnol. Biotechnol. Equip. 31, 630–638.

Liu, L., Tang, L., Tang, M., Zhou, W., 2018. A partially function-to-topic model for protein function prediction. BMC Genomics 19, 883.

Mancano, G., Mora-Ortiz, M., Claus, S.P., 2018. Recent developments in nutrimetabolomics: from food characterisation to disease prevention. Curr. Opin. Food Sci. 22, 145–152.

Mansur, A.R., Seo, D., Song, E., Song, N., Hwang, S.H., Yoo, M., Nam, T.G., 2019. Identifying potential spoilage markers in beef stored in chilled air or vacuum packaging by HS-SPME-GC-TOF/MS coupled with multivariate analysis. LWT Food Sci. Technol. 112, 108256.

Martinović, T., Šrajer Gajdošik, M., Josić, D., 2018. Sample preparation in foodomic analyses. Electrophoresis 39, 1527–1542.

Matsutani, T., Ueno, Y., Fukunaga, T., Hamada, M., 2019. Discovering novel mutation signatures by latent Dirichlet allocation with variational Bayes inference. Bioinformatics 35, 4543–4552.

Miguel, H., Simó, C., García-Cañas, V., Ibáñez, E., Alejandro, C., 2012. Foodomics: MS-based strategies in modern food science and nutrition. Mass Spectrom. Rev. 31, 49–69.

Mikš-Krajnik, M., Yoon, Y., Ukuku, D.O., Yuk, H., 2016. Volatile chemical spoilage indexes of raw Atlantic salmon (*Salmo salar*) stored under aerobic condition in relation to microbiological and sensory shelf lives. Food Microbiol. 53 (Part B), 182–191.

Murzintcev, N., 2019. ldatuning: Tuning of the Latent Dirichlet Allocation Models Parameters. R package version 1.0.0. URL. https://CRAN.R-project.org/package=ldatuning.

Nolasco, D., Oliveira, J., 2019. Subevents detection through topic modeling in social media posts. Future Gener. Comp. Sy. 93, 290–303.

Odeyemi, O.A., Burke, C.M., Bolch, C.C.J., Stanley, R., 2018. Seafood spoilage microbiota and associated volatile organic compounds at different storage temperatures and packaging conditions. Int. J. of Food Microbiol. 280, 87–99.

Park, H., Park, T., Lee, Y., 2019. Partially collapsed Gibbs sampling for latent Dirichlet allocation. Expert Systems with Applications 131, 208–218.

Pavase, T.R., Lin, H., Shaikh, Q., Hussain, S., Li, Z., Ahmed, I., Lv, L., Sun, L., Shah, S.B.H., Kalhoro, M.T., 2018. Recent advances of conjugated polymer (CP) nanocomposite-based chemical sensors and their applications in food spoilage detection: a comprehensive review. Sens. Actuators B Chem. 273, 1113–1138.

Perina, A., Lovato, P., Murino, V., Bicego, M., 2010. Biologically-aware Latent Dirichlet Allocation (BaLDA) for the classification of expression microarray. In: Dijkstra, T.M.H., Tsivtsivadze, E., Marchiori, E., Heskes, T. (Eds.), Pattern Recognition in Bioinformatics. PRIB 2010. Lecture Notes in Computer Science, vol. 6282. Springer, Berlin, Heidelberg.

Pinu, F.R., 2016. Early detection of food pathogens and food spoilage microorganisms: application of metabolomics. Trends Food Sci. Technol. 54, 213–215.

Poghossian, A., Geissler, H., Schöning, M.J., 2019. Rapid methods and sensors for milk quality monitoring and spoilage detection. Biosens. Bioelectron. 140, 111272.

Pratanwanich, N., Lio, P., 2014. Exploring the complexity of pathway–drug relationships using latent Dirichlet allocation. Comput. Biol. Chem. 53, 144–152.

R Core Team, 2019. R: a language and environment for statistical computing. In: R Foundation for Statistical Computing. Austria. URL, Vienna. https://www.R-project.org/.

Rajkovic, A., Smigic, N., Devlieghere, F., 2011. Growth of Escherichia coli O157:H7 and Listeria monocytogenes with prior resistance to intense pulsed light and lactic acid. Food Microbiology 28, 869–872.

Shiraishi, Y., Tremmel, G., Miyano, S., Stephens, M., 2015. A simple model-based approach to inferring and visualizing cancer mutation signatures. PLoS Genet. 11, e1005657.

Silge, J., Robinson, D., 2016. tidytext: text mining and analysis using tidy data principles in R. J. Open Source Softw. 1, 37.

Suzuki, R., Terada, Y., Shimodaira, H., 2019. pvclust: Hierarchical Clustering with P-Values via Multiscale Bootstrap Resampling. R package version 2.2-0. URL. https://CRAN.R-project.org/package=pvclust.

Uyttendaele, M., Rajkovic, A., Van Houteghem, N., Boon, N., Thas, O., Debevere, J., Devlieghere, F., 2008. Multi-method approach indicates no presence of sub-lethally injured Listeria monocytogenes cells after mild heat treatment. International Journal of Food Microbiology 123, 262–268.

van der Hooft, J., Wandy, J., Barrett, M.P., Burgess, K.E.V., Rogers, S., 2016. Topic modeling for untargeted substructure exploration in metabolomics. Proc. Natl. Acad. Sci. 113, 13738–13743.

Wang, Y., Li, Y., Yang, J., Ruan, J., Sun, C., 2016. Microbial volatile organic compounds and their application in microorganism identification in foodstuff. TrAC Trend. Anal. Chem. 78, 1–16.

Warnes, G.R., Bolker, B., Bonebakker, L., Gentleman, R., Liaw, W.H.A., Lumley, T., Maechler, M., Magnusson, A., Moeller, S., Schwartz, M., Venables, B., 2020. gplots: Various R Programming Tools for Plotting Data. R package version 3.0.3. URL. https://CRAN.R-project.org/package=gplots.

Wickham, H., 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York.

Xiong, H., Cheng, Y., Zhao, W., Liu, J., 2019. Analyzing scientific research topics in manufacturing field using a topic model. Comput. Ind. Eng. 135, 333–347.

Xu, Y., 2017. Foodomics: a novel approach for food microbiology. TrAC Trend. Anal. Chem. 96, 14–21.

Yang, M., Qu, Q., Chen, X., Tu, W., Shen, Y., Zhu, J., 2019. Discovering author interest evolution in order-sensitive and semantic-aware topic modeling. Inform. Sciences 486, 271–286.

Yu, K., Gong, B., Lee, M., Liu, Z., Xu, J., Perkins, R., Tong, W., 2014. Discovering functional modules by topic modeling RNA-Seq based toxicogenomic data. Chem. Res. Toxicol. 27, 1528–1536.

Zhang, J., Liu, B., He, J., Ma, L., Li, J., 2012. Inferring functional miRNA–mRNA regulatory modules in epithelial–mesenchymal transition with a probabilistic topic model. Comput. Biol. Med. 42, 428–437.