Detecting Contamination in Shochu Using the Agilent GC/MSD, Mass Profiler Professional, and Sample Class Prediction Models

Application Note

Food Testing & Agriculture

Author

Takeshi Serino
Agilent Technologies, Inc.
Santa Clara, CA
USA

Abstract

Sample Class Prediction (SCP) models are powerful tools that can use mass spectrometry data from highly complex samples to identify differences in sample classes, such as contamination. In this case, a method was developed that uses SCP to accurately detect and classify contamination in shochu samples, for use in quality assurance (QA) and quality control (QC) during the manufacturing process.

Introduction

Shochu is a distilled alcoholic beverage that has been made in Japan since at least the 16th century. It typically contains 25% alcohol by volume and is produced by single or multiple distillation of rice, barley, potatoes or brown sugar. In contrast to wine or traditional liquor, shochu is derived from fermentation by mold instead of yeast. A boom in the consumption of shochu in Japan occurred in the early 2000's, as it became trendy among young drinkers, particularly women. It is perceived to have health benefits such as prevention of thrombosis, heart attacks, and diabetes. As a result, consumption of shochu now exceeds that of sake in Japan.

Quality control for this commercially valuable product is critical to maintain customer satisfaction. Contamination from items such as machine oil and rubber gloves can occur during the manufacturing process. This affects the odor and taste of the product and threatens sales volume for the manufacturer. Timely identification of contaminated product, before it is bottled and shipped, is critical to maintaining brand loyalty.

Agilent Technologies

This application note demonstrates the feasibility of developing a model that can detect the presence of contaminants in shochu during the fermentation and bottling processes. It uses a nontargeted compound analysis approach similar to that recently used for wine classification [1] and determine whether extra virgin olive oil will pass the sensory test [2]. The data were obtained by gas chromatography/mass spectrometry (GC/MS), using the Agilent 7890 GC System equipped with a low thermal mass (LTM) column and coupled to the single-quadrupole Agilent 5975 GC/MS system. However, the total ion current (TIC) traces revealed little difference between most of the sample conditions. Further data processing was required to reveal these differences. The data was then processed using NIST Automated Mass Spectral Deconvolution and Identification Software (AMDIS) and analyzed using a multivariate software package in Mass Profiler Professional (MPP) that includes class prediction algorithms.

Clean shochu samples and samples intentionally contaminated with rubber and machine oil were analyzed, and the data was filtered three different ways to create entities. Sample Class Prediction (SCP) models were then applied to the entities generated by each filter to determine which model was most suited for routine analysis and screening for contaminants. For this data set, the Decision Tree model applied to data filtered using any of the three applied filters provided 100% accuracy in determining the presence of contaminants in samples that were not used to train the model. The Support Vector Machine (SVM) model also provided 100% accuracy, but only when used with data filtered through the Analysis of Variance (ANOVA) plus 45% Coefficient of Variation (CV) filter.

Experimental

Samples

In total, 10 shochu samples were obtained from various sources, some spiked with known amounts of detergents, insecticides, rubber gloves, or machine oil. Samples GA and GB were prepared by adding 20-23 mg of a piece of rubber glove to 1 g of shochu. Sample DA was prepared by adding 10 mg of chlorine detergent to 1 g of shochu, while sample DC was prepared by spraying insecticide detergent (two sprays) into 1 g of shochu. Sample OB was prepared by adding 10 to 15 mg of machine oil to 1 g of shochu. The sample names and their sources are listed in Table 1.

Table 1. Sample Types Analyzed

Sample name	Origin	Contaminants
IOSK	Osaka	None
ITKO	Tokyo	None
IUSA	San Jose	None
DA	Osaka	Chlorine detergent
DC	Osaka	Insecticide detergent
GA	Osaka	Rubber glove A
GB	Osaka	Rubber glove B
OB	Osaka	Machine oil B

Instruments

This study was performed on a 7890 GC System equipped with a low thermal mass (LTM) module and coupled to a 5975 GC/MS system. The instrument conditions are listed in Table 2. The Gerstel MPS2 Autosampler was used to carry out the solid phase microextraction (SPME) sample preparation.

Table 2. GC and Mass Spectrometer Conditions

GC run conditions	
Pre-column	None
Analytical column	10 m × 0.18 mm, 0.18 µm DB-1ms LTM Column Module (p/n 100-2000LTM)
Injection method	SPME (Supelco 57341-U), 1 cm injection
Inlet temperature	Isothermal at 240 °C
Injection mode	1.52 minute splitless at 72 kPa
Oven temperatures	GC oven: 11.83-minute hold at 200 °C (isothermal) LTM module: 120-second hold at 35 °C 35 °C to 240 °C at 30 °C/min Hold at 240 °C for 3 minutes
Column flow	1.1 mL/min constant flow
Carrier gas	Helium
Transfer line temperature	240 °C
GC run time	11.83 minutes
MS conditions	
Ionization mode	EI
Ion source temperature	230 °C
Acquisition mode	Scan (35–450 amu)
Trace ion detection	On
Tuning	atune.u

Sample Preparation

The volatile odor components from each sample type were collected using SPME. Each shochu sample was transferred to 20-mL headspace vials. A 100 µm × 1 cm polydimethylsiloxane (PDMS) SPME fiber (Supelco 57341-U) was exposed to the headspace of the sample at 40 °C for 40 minutes with no agitation. Volatile compounds absorbed on the SPME fiber were thermally desorbed at 240 °C for 1.5 minutes into an injection port. The fiber was baked out in a bake station at 260 °C for 5 minutes after each injection.

Data Processing and Statistical Analysis

Component extraction from the GC/MS data was done using AMDIS on the Agilent MSD Productivity Chemstation (E.02.02). The ELU files from AMDIS were imported into MPP for differential analysis.

Mass Profiler Professional (B.02.02) was used for data filtering and statistical analysis, and Agilent Sample Class Predictor (B.02.) was used to generate sample class prediction (SCP) models. The data processing steps are shown in Figure 1.

- Filter and alignment of compound peaks across samples
- Filtering the entities
- Principal Component Analysis (PCA)
- Hierarchal Cluster Analysis (HCA)
- Create prediction models

Results and Discussion

Data Acquisition

The analysis of shochu samples was performed to survey the compounds that could be detected by GC/MSD (Figure 2). AMDIS was used to extract components from the GC/MS data. The data consisted of four replicates each of rubber, detergent, and machine oil contaminated shochu, 23 replicates of uncontaminated shochu sold in Osaka, 18 replicates of shochu sold in Tokyo, and 13 replicates of shochu sold in San Jose, CA. Typically, 330 to 380 peaks were identified by chromatographic deconvolution. The total ion current (TIC) traces revealed little difference between most of the sample conditions. Further data processing was required to reveal these differences.

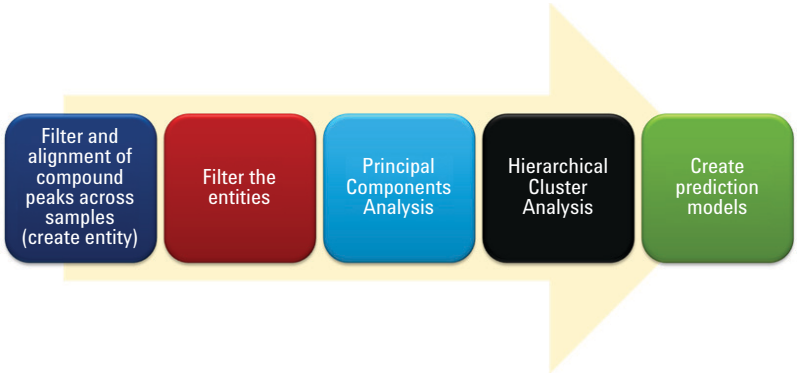

Figure 1. Statistical analysis workflow for generation of predictive models of contamination in shochu from GC/MSD data.

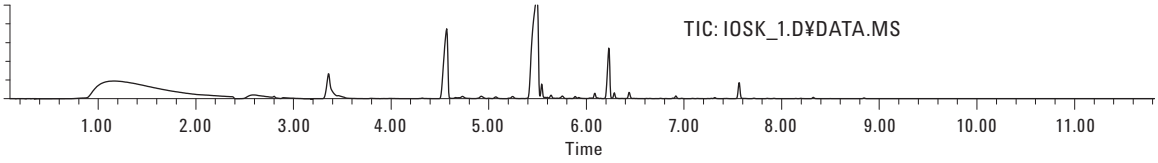

Figure 2. A typical total ion chromatogram (TIC) for analysis of uncontaminated shochu, in this case, sample IOSK.

Data Filtering

As the data set was imported into Mass Profiler Professional (MPP) software, the unidentified components were aligned by spectral similarity and retention time window to form an entity list of 2,376 components. Three entity filters were evaluated to identify entities that could be used to differentiate the various sample types (Figure 3).

The MPP Frequency Table (Filter 1) revealed that many of the compounds were unique to one sample. In fact, 226 entities passed this filter. The one-way analysis of variance (ANOVA) filter (Filter 2) was used with a probability p value of .05 (95% probability that the entity is significant), resulting in 1,080 entities. The third filter selected entities which passed the one-way ANOVA filter and the coefficient of variation (CV) filter set at less than 45% for all samples. The objective of using ANOVA plus CV<45% was to intentionally create a strong filter and investigate its impact on the accuracy of the resulting SCP model.

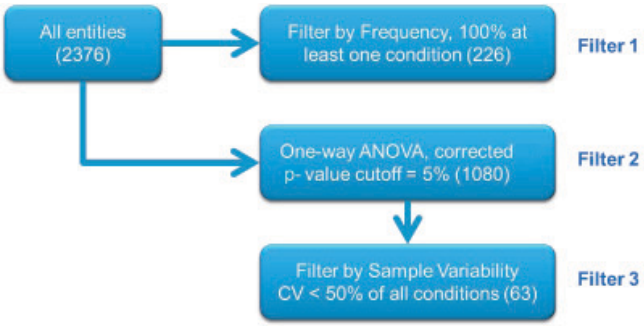

Figure 3. The three data processing filters used to screen entities that were then used to construct SCP models.

Principle Component Analysis

Principal Component Analysis (PCA) was done on the entity lists resulting from the three filters. The PCA score plots are shown in Figure 4, illustrating that the ANOVA + 45% CV filter provides the best separation of all of the sample types, enabling the most accurate SCP models.

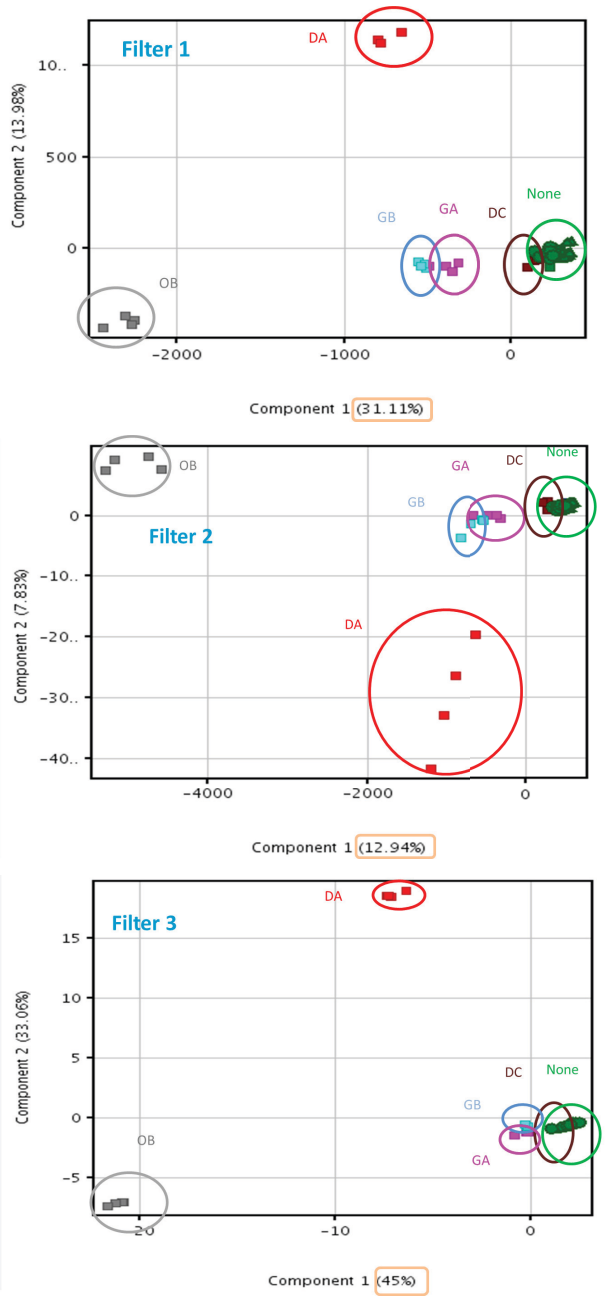

Figure 4. Principal Component Analysis (PCA) scores illustrate that Filter 3 (ANOVA + 45%CV) provides the best separation between the sample types.

Hierarchical Cluster Analysis (HCA)

Cluster analysis is a powerful method to organize compounds or entities and conditions in the dataset into clusters based on the similarity of their abundance profiles. Hierarchical Clustering is one of the simplest and most widely used clustering techniques for analysis of mass abundance data. The method follows an agglomerative approach in which the most similar abundance profiles are joined together to form a group. These are further joined in a tree structure, until all data forms a single group. HCA of the entities generated by Filter 3 found 33 components that distinguished contaminated samples from uncontaminated samples (Figure 5). A library search of these contaminant components against the Wiley 9th/Nist08 library provided identification for eight of them (Table 3).

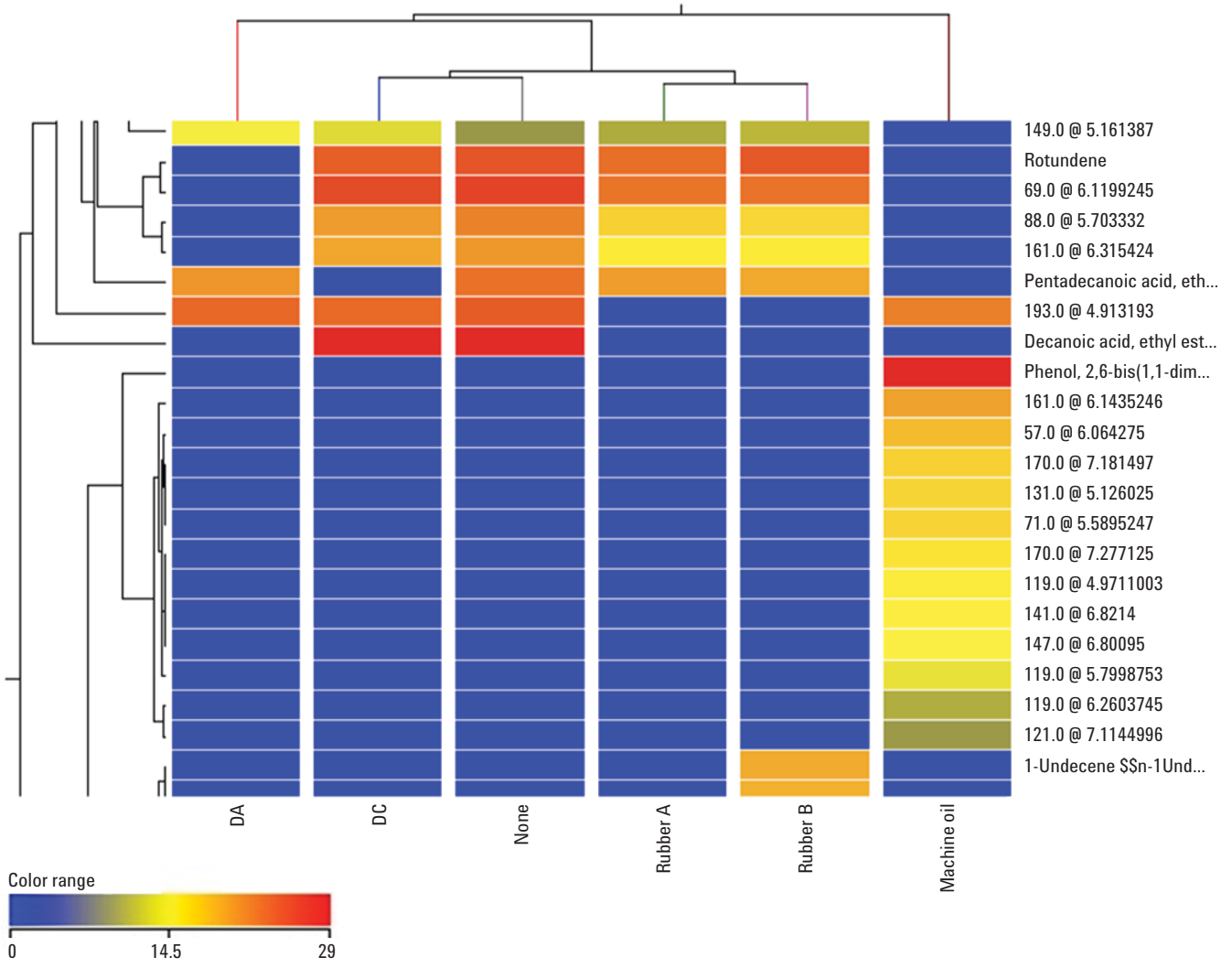

Figure 5. Expanded view of the hierarchical cluster analysis (HCA) heat map for association of compounds detected in the various sample types. The map shows 23 of the 33 components that distinguished contaminated samples from uncontaminated samples. The color range indicates the peak intensity of each compound for each condition: DA; None (uncontaminated machine oil); DC, and so forth. Compounds with low intensity are shown in blue; those with intermediate intensity are shown in yellow, and those with high intensity are shown in red. The color range bar indicates the relative intensity of each compound.

Table 3. Identity of Contaminant Compounds

Sample class	Compound
Machine oil	Phenol, 2,6-bis(1,1-dimethylethyl)-4-methyl- (CAS)
Rubber B	2-Isopropyl-5-methyl-1-heptanol
	1-Undecene
Detergent A	Heptadecanoic acid, ethyl ester
	1-Nonadecene
	1-Decanol (CAS)
Detergent C	Benzamide, N,N-diethyl-3-methyl-
	1,1'-Biphenyl, 3-(1-methylethyl)-

Class Prediction Models

The goal of classification is to produce general hypotheses based on a training set of examples that are described by several variables and identified by known labels corresponding to the class information. The task is to learn the mapping from the former to the latter. Numerous techniques based either on statistics or on artificial intelligence have been developed for that purpose [3]. In this case, the goal was to determine whether a shochu sample was contaminated, based on the 33 compounds that were shown to be associated with contamination.

Five different algorithms - Decision Tree (DT), Support Vector Machines (SVM), Naïve Bayes (NB), Neural Network (NN), and Partial Least Square Discrimination Analysis (PLSDA) - were evaluated to determine which algorithm was best suited for screening for contaminants. Each algorithm was tested with data sets produced using the three data filters.

The Decision Tree algorithm uses a sequence of if-then-otherwise decisions arranged as a tree. A sample gets classified by following the appropriate path down the decision tree. The Support Vector Machines algorithm attempts to separate samples into classes by imagining these to be points in space and then determining a separating plane which separates the two classes of points. The Naive Bayesian classifier assumes that the effect of an attribute on a given class is independent of the value of other attributes. This assumption is called the class conditional independence. A Neural Network is a parallel system inspired by the structure and/or functional aspects of biological neural networks, and it is capable of resolving paradigms that linear computing cannot. Partial Least Square Discrimination Analysis is particularly adapted to situations where there are fewer observations than measured variables. It is used to sharpen the partition between groups of observations, such that a maximum separation among classes is obtained.

The first step in building the classification model was to train the models with the data, using each of the five model algorithms with each of the three data filters. To validate each model, the same training data were used. Although redundant, this is a valid statistical procedure. The prediction accuracy of each the five models with each of the data filters for the training data is shown in Table 4. The Support Vector Model and Decision Tree algorithms were able to validate the models to 100% accuracy using all three filters.

Table 4. SCP Model Accuracy (%) After Training

Algorithm	Filter 1	Filter 2	Filter 3
Decision tree	100%	100%	100%
Naïve bayes	99	100	99
Neural network	69	73	69
Partial least square Discriminant analysis	90	90	90
Support vector model	100	100	100

The second step was to test each model with unknown sample data. An additional 12 samples that were not used to create the models were used for this purpose (Table 5). Using these samples, the DT model is more robust when predicting unknown contamination, since it provides 100% accuracy with all three filters (Table 5). In contrast, the SVM model provides 100% accuracy only with the ANOVA + 45% CV filter (Filter 3). This implies that there is a limited set of entities that heavily influence the prediction model. The PLSDA model did a very poor job of identifying contaminated samples. The variation in behavior of the modeling algorithms illustrates the need to choose the best model for a given set of data.

Table 5. SCP Model Accuracy (%) for Unknown Samples

Algorithm	Filter 1	Filter 2	Filter 3
Decision tree	100%	100%	100%
Naïve bayes	92	100	92
Neural network	50	50	50
Partial least square discriminant analysis	50	42	33
Support vector model	83	83	100

Conclusions

Sample Class Prediction (SCP) provides a robust way to determine shochu quality that can be used in a production QC environment. Small differences between samples can be clearly seen, using a multivariate analysis of GC/MSD data.

In order to generate the SCP model with the highest accuracy of prediction, the data quality is crucial. This facilitates construction of the right filtering and prediction model for the samples. SCP will provide the best results when the sample data is properly filtered and the proper prediction algorithm is used. Multiple prediction models allow the evaluation and customization of different prediction models to the analysis. Better entity lists enable the development of better SCP prediction models which in turn enable improvement of the workflow of QA and QC of food analysis.

References

1. L. Vaclavik, O. Lacina, J. Hajslova, J. Zweigenbaum. "The use of high performance liquid chromatography-quadrupole time-of-flight mass spectrometry coupled to advanced data mining and chemometric tools for discrimination and classification of red wines according to their variety." *Anal Chim Acta*. **685**, 45-51 (2011).

2. S. Baumann and S. Aronova. "Olive Oil Characterization using Agilent GC/Q-TOF MS and Mass Profiler Professional Software" Agilent Technologies Application Note 5991-0106EN.

3. J. Boccard, J. L. Veuthey, S. Rudaz. "Knowledge discovery in metabolomics: an overview of MS data handling." *J Sep Sci.* **33**, 290-304 (2010).

For More Information

For more information on our products and services, visit our Web site at www.agilent.com/chem.

www.agilent.com/chem

Printed in the USA
August 2, 2012
5991-0975EN

Agilent Technologies