# Cannabis Chemometrics: Converting Untargeted Data to Knowledge

By Cindy S. Orser Ph.D., CLIP Labs

Chemometrics is a relatively new field of science that coalesced in the early 1970s when computers were being increasingly used for scientific data analyses. A Swedish scientist, Professor Svante Wold at Umea University, coined the term “chemometrics” in a grant application in 1971 and, thereafter, formed the International Chemometrics Society with the contemporaneous chemometrics pioneer, Professor Bruce Kowalski at the University of Washington. [1]

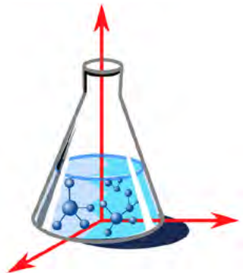

**Figure 1.** The Milano Chemometrics and QSAR Research Group, Department of Earth and Environmental Sciences, University of Milano-Bicocca. https://michem.unimib.it; image used with permission from Professor Roberto Todeschini.

Chemometrics naturally evolved as a multi-disciplinary solution to the complex problem of interpreting and modeling vast multivariate data (simultaneous observation of more than one characteristic).

The advantage brought by applying chemometrics is the analysis of variance and dependence between variables, thereby revealing the contribution of heretofore hidden variables within the spectrum, and incorporating that information into predictive outcomes.

Chemometrics has emerged as a powerful applied science to reveal meaning in the living world through data-analytic disciplines of computer science and applied statistical mathematics. Chemometric generated models are applicable to data obtained from various analytical instruments and influence advancements in methodologies behind state-of-the-art analytical instrumentation.

**Calibration & Classification**

Chemometrics essentially removes noise from multivariate data, identifies latent variables, and provides a means to visualize hidden correlations. Data can either be used for targeted or non-targeted analyses. Targeted analysis (when you are looking for a known analyte) represents the conventional approach, while non-targeted analysis covers the chemical analysis of the entire matrix (e.g., olive oil) to develop a fingerprint for authenticity. The data are analyzed by chemometrics software to extract relevant information through sophisticated statistical mathematics and relating that information back to the chemical process of interest to reveal knowledge about the system or process. This newly found knowledge can then be modeled and applied to make decisions in the future.

Calibration is a central task in chemistry that enables the prediction of properties of interest based on what is being measured (e.g., concentration over a continuous scale). Expanding to a multivariate calibration using chemometric techniques to model data also requires a set of reference values for the properties of interest. For example, to develop a near-infrared (NIR) handheld device that can predict tetrahydrocannabinolic acid (THCA) content in previously unanalyzed, cured cannabis flower, the reference data would take the form of NIR spectra (using the same device) for an analyzed cannabis flower sample of known THCA concentration as previously determined by a validated method such as high-performance liquid chromatography (HPLC). Through this multivariate calibration process, non-destructive analytical measurements of future samples can be taken inexpensively to predict the property of interest (i.e., THCA against an immense background of interfering analytes).

**Data Analysis**

Multivariate data analysis is particularly interesting and well suited for exploratory analyses to interpret patterns in data and develop models. Multivariate classification is a direct application of multivariate calibration in that the training dataset is used to build a model, which can classify the analysis of future samples using the analytical method that was modeled. These models can be routinely applied to future data to predict the same parameters of interest, whether it is to discriminate through targeted analyses of a specific ingredient or to detect adulteration using non-targeted analysis.
Data measurements come in many different forms. Chemometrics has been applied the most using chemical measurements from a variety of analytical instruments, including spectroscopy (mass spectrometry, nuclear magnetic resonance, NIR, Raman, and fluorescence), chromatographic (HPLC), physical (concentration, temperature, pressure, viscosity, and flow rate), and a myriad of molecular methods.

The actual data analysis can be preceded with a preprocessing step where inherent variation in the dataset that does not reflect the analytical method is removed through one of several means. Preprocessing can be accomplished through simple normalization of the data, baseline correction, mean centering, or more sophisticated manipulations that include removing true outliers.

**Basic Modeling**

Data analysis usually involves applying algorithms in the form of principal component analysis (PCA) or Partial Least Squares (PLS). PCA is normally used to explore classification of data by extracting principal components, or sources of variation, in order of their significance in contributing to variation within the data. PCA is also the simplest means to model information from multiple variables (e.g., spectral scans across multiple wavelengths) into synthetic variables (principal components) that summarize and encompass the variation and explain a certain percentage of the observed patterns. PCA is often combined with cluster analysis (CA), another basic modeling tool, to separate samples into groups that share a common property or set of variables. The intra-sample variation is displayed by plotting the resulting scores from Principal Component 1 (PC1) against Principal Component 2 (PC2) which can reveal outliers within the data set that can be removed.

These basic chemometric modeling tools are commonly applied as a quality control for foods like herbs to assure authenticity, as fraud within high-cost foodstuffs is rampant. [2] Within the cannabis industry, PCA and CA have been applied to introduce an alternative to the classification of chemovars. [3,4,5] In a previously published large-scale

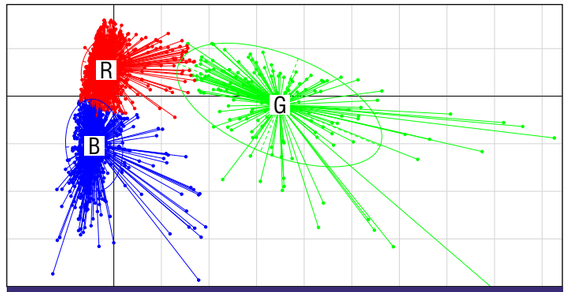


**Figure 2.** PCA and Cluster Analysis of Terpenoid Profiles From 2,237 Individual Cannabis Flower Samples. [3]

Nevada PCA study, 2,237 individual cannabis flower samples representing 204 cultivars from 27 cultivators were analyzed for 11 cannabinoids and 19 terpenoids. [3] Even though 98.3% of the samples were drug type I (high THC), PCA analysis of the combined dataset resulted in three clusters distinguishable by terpene profiles alone (Figure 2), where cannabinoid content was of no distinguishing value.

**Advanced Modeling**

Once a model has been derived, it must be maintained through continual collection of data to re-validate the model. The quality of the model determines the quality of the predictions of future samples. These popular methods have been applied to clinical data for disease diagnosis [6] and for identification of the origin of consumer goods. [7]

**Real World Chemometrics**

Fraud within the food industry is rampant, with global estimated losses at $49 billion USD. [8] The use of multivariate non-targeted spectral analysis is a perfect solution to not only find adulterants in food products but also to confirm authenticity (is a red wine a pinot noir from Northern Italy, New Zealand, or Oregon?). Turns out that this burning question can actually be answered through a chemometrically-derived classification application called WineScreener™ created by the Bruker Corporation, the company whose name is associated with $^{1}$H NMR spectroscopy. This commercial classification analysis tool enables comparison of non-targeted spectral $^{1}$H NMR data with corresponding reference spectra obtained from production sites around the world using verification models. [9] WineScreener is a stepwise process where samples are collected, prepped, and analyzed locally; however, the $^{1}$H NMR data are evaluated on a centralized Bruker server that reports back to the customer.

Food authenticity was first introduced in 1990 in the Food Safety Act to ensure that product is as advertised and free of adulteration. [7] Questions surrounding food authenticity cannot be answered without the existence of robust analytical reference data from authentic samples in the form of chemoprofiles or fingerprints. Chemometrically-derived models developed through the steps discussed above provide the ideal solution for authentication of a given plant material, active ingredient, or common adulterant using non-destructive, rapid, accurate analysis such as NIR or NMR answering the question of whether the material is authentic or fraudulent.

The public health threat from plant material fraud touches many industries, including medicinal herbs, spices, pharmaceuticals, dietary supplements, and cannabis. The concept of equivalence in herbal formulations started in Germany to establish clinically proven reference materials. [10] Ground plant material is difficult if not impossible to taxonomically identify.

And morphological data alone are insufficient when a plant's metabolic chemoprofile is required for authentication and traditional chemical analysis is too time consuming and expensive. High value spices and herbs are subject to known adulterating substitutes (e.g., saffron is often cut with ground turmeric rhizomes or *Cannabis sativa* stamens and black pepper with papaya or chili seeds). [11] Taking a chemometric approach to complex multivariate data can result in a useful tool for authentication.

Public concern over food safety is driving governmental agencies to confront widespread adulteration of food. Significant challenges still exist in the effort to target food fraud, including lack of guidelines regulating the development and validation requirements of non-targeted methods and interpretation of their results, existence of standardized and certified reference materials to build models around, and lack of validated chemometric software. [12]

**The Data Explosion**

A new era of big data is emerging. As chemometric studies have advanced to incorporate many different levels of data including genomic, proteomic, metabolomics, and now metabonomic, the level of sophistication in study design and demand for quality and accuracy in the datasets has intensified. The need for machine learning algorithms has reinvigorated entire academic fields such as quantitative structure-activity relationships (QSARs) and quantitative structure-property relationships (QSPRs) to correlate the molecular structure of chemicals to measurable properties with molecular chemometrics as the basic tool in a highly interdisciplinary environment applying modeling and data analysis to molecular data. [13]



The question to be addressed changes from "what is there" to "what is the relation to," or "what is the difference between?" [14]

**Chemometrics & Cannabis**

Applying chemometric tools to authentication and quality control of cannabis flower and cannabis-based drugs will be both efficient and empowering. One Canadian company, PURITY IQ, in Mississauga, Ontario, recently launched their Global Cannabis Registry as a component of its voluntary good manufacturing practices compliance requirements for third-party certification in the Cannabis Authenticity and Purity Standards (CAPS) and certification program. [15] Purity IQ strives to fill the void between current state government basic safety regulations for cannabis and cannabis-based products and more rigorous safety and quality brand protections in place in other established industry sectors, brands, and retailers.

Applying scientifically validated and chemometrically modeled non-targeted fingerprint data from both genomic and NMR data in partnership with key strategic industry partners, Purity IQ plans to register cannabis cultivars from legal producers with unique fingerprints for easy reference by other parties.

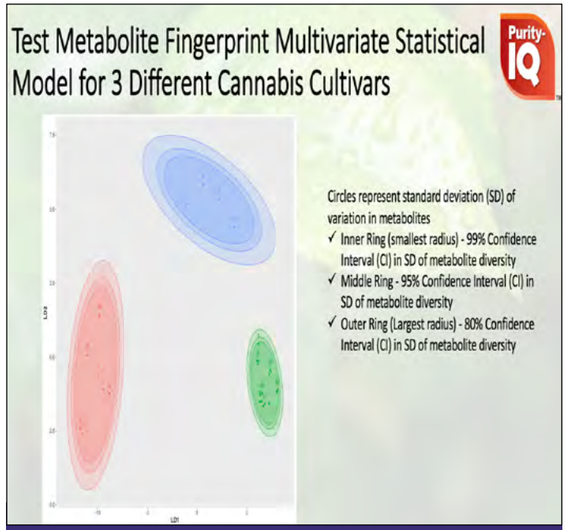


**Figure 3.** Chemometrics on three different cannabis cultivars for multiple batches. Consistency in metabolite profiles is defined by the concentric rings with the inner ring of 98% CI (Confidence Interval) in variation of metabolites; the second ring is 95% CI and the outer ring is 90% CI. Not all cannabis grow operations have such tight controls on consistency of product; some are very inconsistent with highly variable metabolite profiles from batch-to-batch.

Figure 3 demonstrates how a chemometric approach to metabolic data can provide a quick quality check from batch-to-batch production using a specific cannabis cultivar that was included in a validated model.

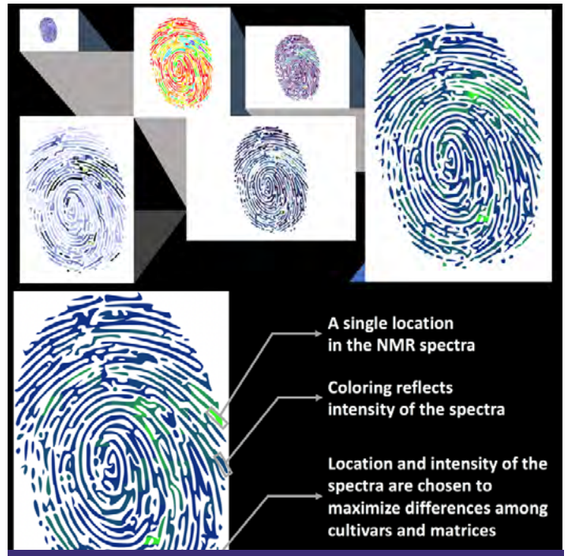


**Figure 4.** A) Chemometric metabolite profiles can be visualized as fingerprint images that are scanned by readers and uniquely identify cannabis products. B) Multiple fingerprints for different cannabis cultivars representing variation in metabolites.

Furthermore, the individual chemometric profiles for a given cannabis sample or cannabis-based product can be represented as a fingerprint that can be scanned to provide the metabolic details for that cultivar's unique fingerprint. (Figure 4)

**Looking Ahead**

The cannabis industry should anticipate widespread non-targeted application of chemometrics increasing in the years ahead with a focus on emerging real-world problems including quality control and fraud. A key to the success of its implementation will be the availability of data from samples used for modeling, external validation set data, as well as external raw data. To date, within the food industry, external raw data are rarely made available.

Therefore, non-targeted screening technologies are currently access-limited due to the proprietary nature of samples stored within reference databases, as within the wine industry. A more practical solution would be the creation of open access databases as has been done for cannabis genomics data, so that the full utility of non-targeted screening can be made widely available.