INSTEP™

AUTOMATION OF CHEMOMETRICS PREDICTIONS

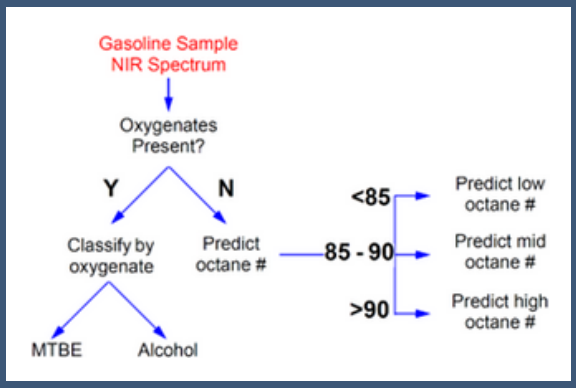
BY INFOMETRIX, INC.

INSTEP™

Integrating and Automating Machine Learning and Chemometrics Models

Just building a model is not the end; InStep finishes the task. The end goal is to use the knowledge you have gained using multivariate analysis to automate the assessment of a continuing stream of data. That is where InStep comes into play. You use InStep to create a custom expert system that will automatically interpret newly acquired data.

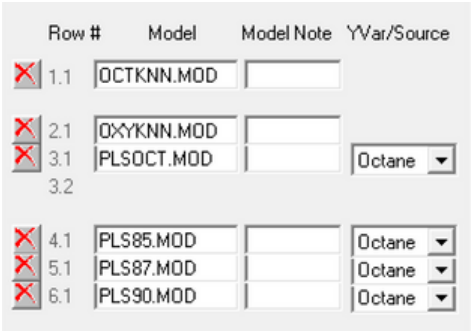
Today's fast processors and large storage devices enable the acquisition of copious amounts of analytical data. Software tools like Pirouette® can detect the presence of meaningful information in that data and create multivariate models able to make predictions about



A decision tree for customized interpretation of gasoline mixing classification and regression models into an expert system

future samples. Automating the characterization of these future samples is the goal of InStep. It combines a decision tree approach with the multivariate data processing intrinsic to Pirouette to facilitate routine prediction. And because the time required to move from R&D to practical use is measured in hours or days, not weeks or months, InStep fills a niche for both laboratory-based and in-line process monitoring and control functions.

Complex applications benefit from the increased prediction accuracy and sensitivity to outliers (e.g., unusual samples, process upsets) associated with a multivariate approach. Generating a decision tree around a classification model (predicting a sample category), a regression model (predicting a continuous property or concentration), or any combination of both model types allows predefined rules to be applied to complex problems. When coupled to a given instrument system, the result is a turn-key custom analyzer.

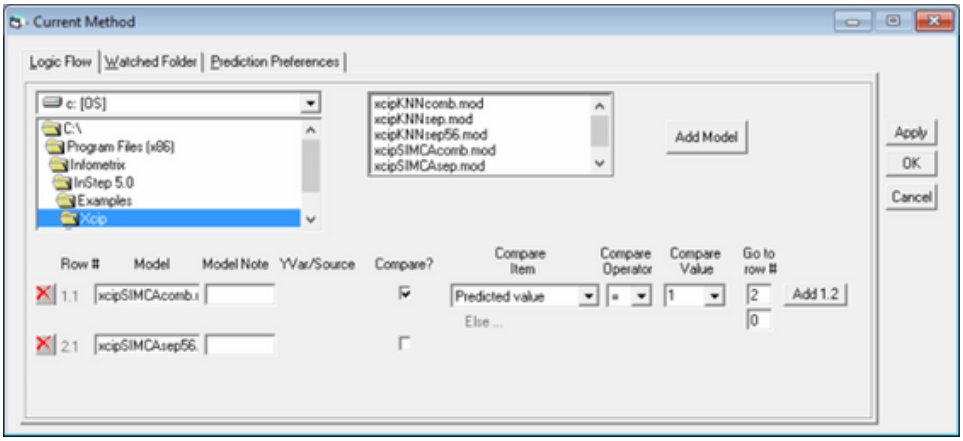


The key aspects of InStep are:

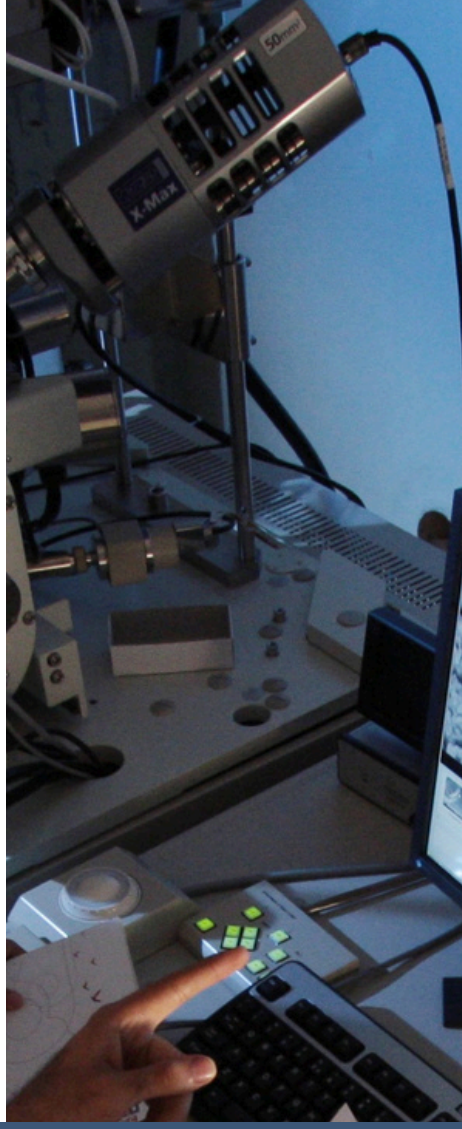
- its reliance on the same core algorithms of Pirouette to make predictions
- a form-based interface for creating decision trees using multivariate models
- a "watched folder" approach which allows it to run concurrently with instrument data acquisition programs
- generation of custom reports which can be displayed on-screen and/or saved to a file·presentation of control charts for visual monitoring of results and diagnostics

InStep Methods and Format

There are two forms to fill out to customize InStep into an application-specific analysis engine. First is to define the pattern you want InStep to follow in processing data. As shown, you can choose one or more models and arrange them in a sequence to interpret data automatically.



With a method in place, designing the report is enabled through the Format menu. Myriad options are available to provide written reports on a sample-by-sample or batch-by-batch basis.

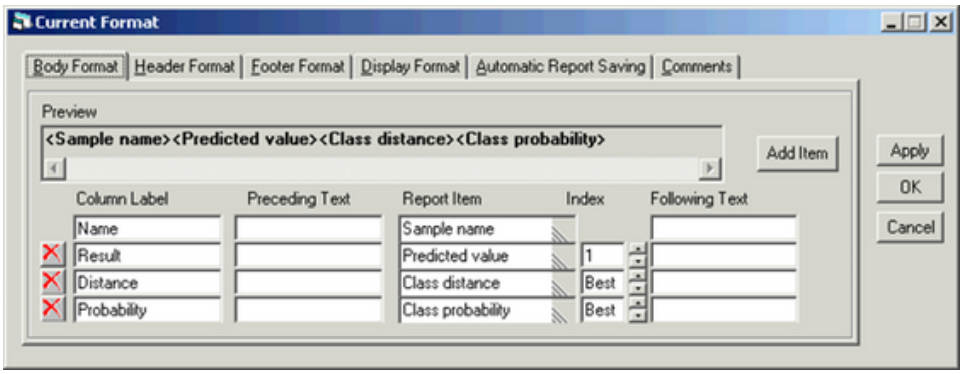


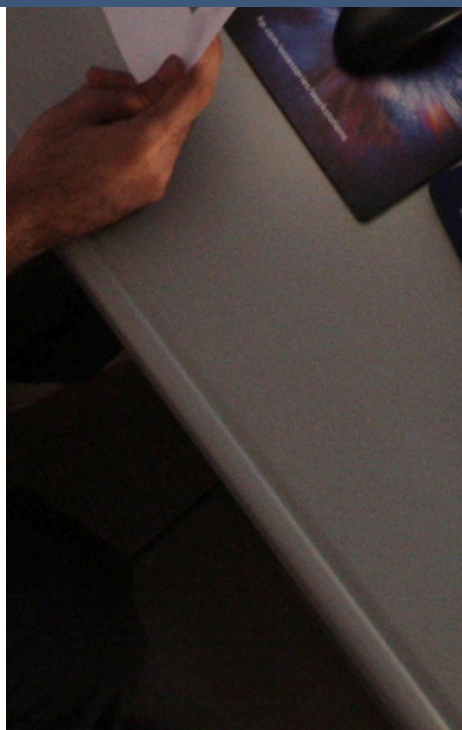
Objective Assessments

Manual interpretation is inefficient and introduces a possible source of error. InStep ensures consistent quality in quality control.

Automation

InStep performs the interpretation of complex patterns in a fully-automated manner and in real time.





Sample Processing

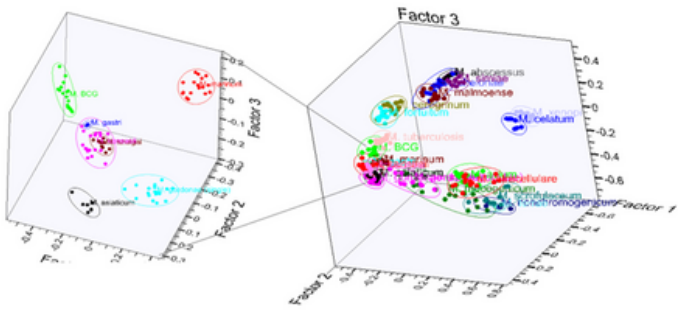
Once a method and format have been chosen, sample processing can begin. Most users employ a watched folder approach. When a file of extension specified by the method appears in the watched folder, it is loaded and the current method is applied to it. A prediction is made using the first model in the method. The prediction results then filter through any comparison logic for that model where branching to other models may occur.

For multi-model methods, a sample may experience several predictions before a final result is reached. At that point a line in the report is generated; its appearance is governed by the current format. If the file contains more than one sample, processing of the next sample then begins. After processing, the file is moved and monitoring of the input folder resumes.

Users may invoke InStep via a command line call which specifies a method, format, and target data file. In this mode InStep processes the target data and then quits.

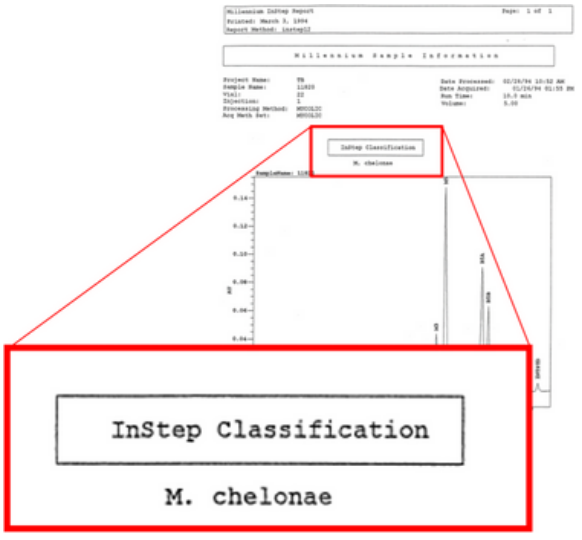
Example 1: Tuberculosis Analysis

A system was put into place for the Centers for Disease Control to automate the interpretation of chromatograms derived from the cells of Mycobacteria, a family of over 60 species related to the dominant tuberculosis. Handling the data in a stepwise fashion, InStep was able to interpret the results of the analysis using a hierarchical cascade of models. Completely automated, the complexity of the analysis was bundled into a single method/format pair.



CDC Model for M. *tuberculosis* and family

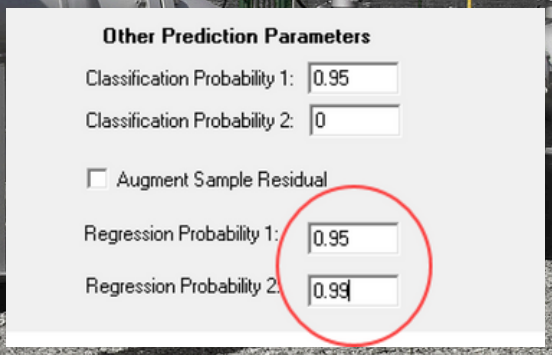
For output, the reported bacterial ID could be printed out in real time with the chromatogram as demonstrated here.



Adding Control Limits

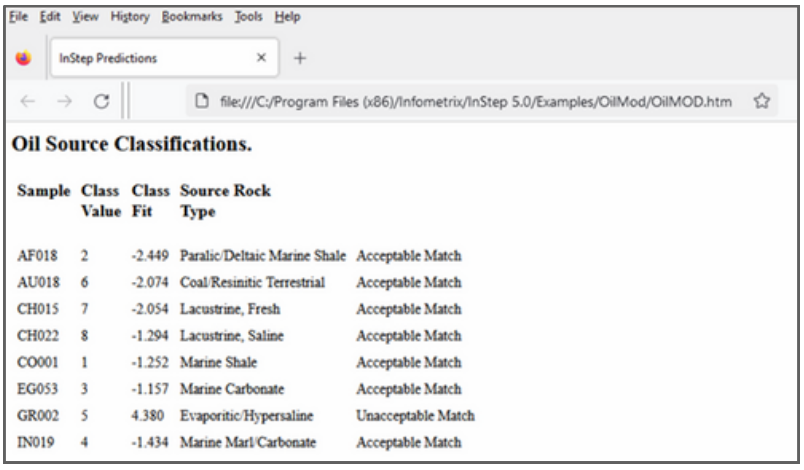
As has been shown, making routine predictions with InStep is quite straightforward. Suppose the application includes a data historian to capture not only the predicted values and outlier diagnostics but requires information from the training set regarding working limits on the predictions.

InStep can retrieve, from the Pirouette model, limits on the Y predictions and outlier diagnostics at levels of confidence specified in the InStep method. An example where the method includes two confidence levels is shown here.



Example 2: Classification of Crude Oils

An example using control limits combined with the power of the format function is shown in the OilMOD application by GeoMark Research. In this case, GCMS data for crude oil samples are processed to identify the type of source rock that was responsible for the sample collected. A Pirouette classification model was built, and two control limits set. The report replaces the control limit screen with words reflecting the quality of the match.



Oil Source Classifications.

Sample	Class Value	Class Fit	Source Rock Type	
AF018	2	-2.449	Paralic/Deltaic Marine Shale	Acceptable Match
AU018	6	-2.074	Coal/Resinitic Terrestrial	Acceptable Match
CH015	7	-2.054	Lacustrine, Fresh	Acceptable Match
CH022	8	-1.294	Lacustrine, Saline	Acceptable Match
CO001	1	-1.252	Marine Shale	Acceptable Match
EG053	3	-1.157	Marine Carbonate	Acceptable Match
GR002	5	4.380	Evaporitic/Hypersaline	Unacceptable Match
IN019	4	-1.434	Marine Marl/Carbonate	Acceptable Match

Transfer of Calibration

It is not unusual for multivariate models to age with changes in the data source. The desirable solution for such a situation would be to recollect the data used to create the original training set, then make a new model to be used in the InStep method. However, collecting a new, full set of training data may be costly. If the original model could be modified to work with the new data, that cost is eliminated. InStep has the ability to perform calibration transfer for these cases.

Automating Interpretation

If you work with chemometrics or machine learning algorithms as applied to analytical or process data sources, the benefit to your organization is to put this knowledge to us in routine assessments.That is the role of InStep.Contact us with any questions on how to deploy these automated interpretation engines.

InfoMetrix®

THE PREMIER CHEMOMETRICS COMPANY

CONTACT US

For evaluation, quotation, presentation or additional information

Infometrix, Inc.
19119 North Creek Pkwy
Suite 100
Bothell, WA 98011

T: 425-402-1450
www.infometrix.com
info@infometrix.com