Viscovery® SOMine 7 – Data Sheet

Viscovery SOMine | Data Mining Suite

Workflow-oriented software suite based on self-organizing maps (SOM) and multivariate statistics for explorative data mining and predictive modeling

General Features of the System

Project environment

- Analytical work organized in projects, providing focused navigation through the application
- Goal-oriented operation, shielding the user from the technology core and the statistical algorithms
- Projects, consisting of up to 5 workflows, provided for data import and preprocessing, SOM creation and exploration, classification, local SOM prediction, and scoring
- Project workflows, optionally processed with minimal user interaction
- Administration of projects, performed in directories with “clean directories” function

Workflow orientation

- Dedicated workflows, each consisting of 4 steps, with clear tasks
- Steps providing proven default settings
- Workflow branching allowing generation of model variations
- Functions allowing parameterizations to be copied between workflow steps
- Workflow branch sorting with drag-and-drop function

Reporting and documentation

- Instant, on-demand reports for each workflow step as well as the entire workflow
- Dynamic reporting of setting differences between steps in a branching
- Integrated documentation for projects
- Production journal of created models
- Functions for adding descriptions and detailed comments to attributes, clusters, and models

Usability

- Informative pop-ups throughout the system, as well as supportive context menus in numerous places
- Multiple handling tools, such as quick search for strings and sortable tables
- Transfer of attribute selections and retention of attribute order between lists with copy and paste
- Function for saving combinations of visible windows with selected properties as “Arrangements”
- Multiple options for setting preferences for charts, tables, statistics, modeling, and visualization
- Online help function and comprehensive user manual

Proprietary Data Formats

Viscovery projects

- *.visdm files: proprietary format for Viscovery data mining project files, which contain all relevant information of a Viscovery SOMine analysis project, including all preprocessing and other settings
- Earlier Viscovery project file formats (*.vsp, *.csp, and *.vpp from version 4, 5, and 6 of Viscovery SOMine, Viscovery Profiler, and Viscovery Predictor) can be imported

Viscovery® SOMine 7 – Data Sheet

Viscovery data marts

- *.dms files: Viscovery data mart files, containing preprocessing and other meta information
- *.dmd files: associated Viscovery data mart files, containing portions of imported data, optimized for rapid access

Viscovery models

- *.som files: Viscovery SOMine model files, which contain the SOM data representation, including preprocessing settings, formulas, clusters, and predictive modeling definitions
- Older Viscovery model file formats (*.som of versions 4, 5, and 6 of Viscovery SOMine, Viscovery Profiler, and Viscovery Predictor) can be imported

Technical Requirements and Configuration

Minimum workstation configuration (recommended configuration)

- 2 GHz CPU (3 GHz or higher)
- 2 GB RAM (8 GB or more)
- Windows 7 or newer
- 32-bit OS (64-bit for voluminous data sets)
- 24 bit color graphics, 1280 x 800 (full HD or higher)
- Required disk space is 6 times the size of the analyzed data file (10 times for comfortable project work)

Modular software configuration

- The minimum configuration of Viscovery SOMine consists of the **Visual Clusters** core module.
- For advanced exploration and modeling, the **Explore and Classify** and **Predict and Score** extension modules are available.
- The **Enterprise Data** extension module provides enhanced features for data import and handling, as well as for the analyses of very high-dimensional and voluminous data.
- Each workflow can be automated using the corresponding **Workflow Automation Service**, which is available as part of an additional package.

Licensing

- Each modular configuration is available with a perpetual license or, alternatively, with a term-license for a specified period of time.
- Each modular configuration is available as a single-user license or as a network license.
- Single-user licenses are bound to the user account on a single computer and may be transferred to a different computer system once per year.
- Network licenses allow operation of the software for a defined maximum of concurrent users and require the additional installation of the Viscovery License Server.
- Each licensed user may operate one interactive instance of the software at a time.
- All configuration and validity information regarding licenses is coded in unique license keys.
- Software installation and license activation require administrator privileges and Internet connectivity.

Viscovery® SOMine 7 – Data Sheet

Basic Module | Visual Clusters

Visual data mining with self-organizing maps

Main Functions and Features

General characteristics

- **Visual Clusters** is the core module of the Viscovery SOMine Suite. It can be operated as a stand-alone system, as well as be flexibly extended with other modules of the suite.
- The two basic **Preprocess** and **Cluster** workflows are available; the first guiding the user through data import and preprocessing, the latter through the creation and exploration of the SOM model.
- Data sets with up to 100,000 records and up to 100 variables can be processed.

Data preprocessing

- Determination of variable names, types, and descriptive statistics
- Optional renaming of attributes
- Cross-reference definition from data records to external documents (links)
- Definition of new variables using built-in formula language
- Definition and automatic management of nominal variables (single-valued and multi-valued)
- Transformation of variables
- Treatment of outliers
- Replacement specification for ranges and special or missing values
- Conditional removal of data records
- Statistical and deterministic sampling and over-sampling of records

Data representation through self-organizing maps (SOM)

- High-performance computation of batch SOM based on classical Kohonen algorithm
- Two-dimensional SOM data representation on a hexagonal grid
- Predefined training schedules with selectable map size, granularity, and tension
- Automatic standardization of variables with additional scaling options
- Definition of the influences from individual attributes on the map ordering by setting attribute priorities
- Automatic compensation of correlations in the data
- Well defined treatment of missing values in all stages of model creation
- Optional setting of default parameters for map creation

SOM visualization and exploration

- Interactive visualization of attribute distributions and characteristic values in a map window
- Automatic color-coding of attributes with transformation-adjusted color scale or black-and-white options
- Annotation of the map with labels
- Manual drawing of trajectories and selections in the map
- Various options for selecting and unselecting map regions (by color-scale, interval, cluster, etc.)
- Display of thumbnails from external documents over the map window
- Display of nearest neighbors from the active node in the map
- Sorting of attributes in the map window according to similarity or priority

Viscovery® SOMine 7 – Data Sheet

Visual cluster analysis

- Automatic execution of agglomerative clustering methods (SOM-Ward, Ward, SOM-Single-Linkage)
- Selection of preferred cluster method and initial number of clusters prior to map creation
- Integrated visualization of cluster boundaries, cluster centers, and inner structures
- Display of separate clusters with optional color-coding of clusters (flat, shaded, or U-matrix)
- Display of cluster means for all attributes in the statistics pane
- Creation of map labels from cluster names

Statistical analysis of data associated with workflow steps

- Descriptive statistics
- Attribute histograms
- Correlation analysis
- Principal component analysis
- Frequency tables
- Box plots
- Scatter plots

Access to original data

- Data-record browser for showing original data from any active workflow step or selected region in the map
- Opening of external documents by clicking a region of the map
- Import and export of map labels, selections, and paths from/to external files

Available data interfaces

- Import and export of tab-delimited flat-text files (*.txt) and Microsoft Excel files (*.xlsx)
- Import of space-separated flat-text files (*.txt) and Excel 97/2000/2003 files (*.xls)
- Intelligent copy-and-paste function between Viscovery internal data and external software
- Export of SOM node values as a table

Viscovery® SOMine 7 – Data Sheet

Extension Module | Explore and Classify

Advanced SOM model exploration, clustering, and classification

Main Functions and Features

General characteristics

- **Explore and Classify** is an extension module of the Viscovery SOMine Suite. It requires the **Visual Clusters** core module and can optionally be combined with other modules of the suite.
- The module provides multiple features for advanced exploration of the self-organizing map (SOM) model and for the interactive definition of alternative segmentations. In addition, the **Classify** workflow guides the user through the application of segmentations to new data and the evaluation of classification results in comparison to control groups.

Group profiling and comparison

- Identification of significant variable deviations using interactive profiling of data groups (pools)
- Creation of groups from clusters, nodes, neighborhoods, or arbitrary selections
- Evaluation of differences between two arbitrary groups using a statistical contrast
- Define reference group from any cluster, node, neighborhood, or arbitrary selection
- Evaluation of non-trivial group descriptives using the “Profile” chart and “Cluster” pane

Interactive data statistics for arbitrary map regions

- Descriptive statistics
- Attribute histograms
- Correlation analysis
- Principal component analysis
- Frequency tables
- Box plots
- Scatter plots

Cluster characteristics

- Automated comparison of existing clusters regarding significant differences in attribute means
- Display of attributes characterizing each cluster based on cluster profiles
- Option for showing each attribute in the cluster where it is extreme

System state monitoring

- Dynamic simulation of process trajectories in the SOM model
- Sequential reading of records from time-ordered data file
- Selection of velocity and trace length of simulation

Multiple segmentations

- Creation of additional segmentations in an existing SOM model
- Interactive selection of cluster method, number of clusters, and attributes for new segmentation
- Display of existing segmentations in the dedicated “Segmentation” window
- Annotation, comparison, copy, and deletion of segmentations

Viscovery® SOMine 7 – Data Sheet

Interactive cluster definition

- Manual adjustment of cluster boundaries by joining and deleting clusters
- Definition of a new cluster or extension of an existing cluster with an arbitrary selected area
- Automatic determination of cluster names from a nominal attribute
- Annotation and renaming of clusters
- Assignment of post-processing formulas to clusters

Classification

- Application of models to new data with the dedicated **Classify** workflow
- Selection of a segmentation as classifier
- Automatic random generation and administration of control groups
- Rapid classification of data records of a new data mart, including evaluation of post-processing formulas
- Export of classification results and, for each record, associated cluster statistics and map node values to tab-delimited flat-text files (*.txt) and Microsoft Excel files (*.xlsx)

Evaluation of classification results

- Visualization of applied data distributions over the SOM model
- Specification of real class names in an additional attribute for test or evaluation purposes
- Additional visualization of classification error over the map if real class memberships are known
- Generation of charts for evaluation of the model application and comparison to the control group

Viscovery® SOMine 7 – Data Sheet

Extension Module | Predict and Score

Non-linear predictive models based on self-organizing maps

Main Functions and Features

General characteristics

- **Predict and Score** is an extension module of the Viscovery SOMine Suite. It requires the **Visual Clusters** core module and can optionally be combined with other modules of the suite.
- The module provides two additional workflows: the **Predict** workflow guides the user through the creation and validation of linear and non-linear prediction models, the **Score** workflow through the definition and application of scoring models to new data and through the evaluation of scoring results in comparison to control groups.

Linear regression models

- Deterministic and random partitioning of data into a training set and up to two test data sets
- Computation of global models as multi-linear or, optionally, stepwise linear regressions
- Option for logistic approximation of binary target variables
- Model statistics for resulting regression and common estimates for determination coefficient, standard error, and prediction interval for selected confidence level
- Beta coefficients for independent variables, including t-statistics, tolerance, and “linear influence” on the target variable
- Storage of linear regression models as PMML code in *.vxm auxiliary files, which can be directly imported into PMML-supporting engines from other vendors

Local regressions and non-linearity analysis

- Based on a patented procedure that combines self-organizing map (SOM) technology with statistical “white-box” models, designed for explaining residual variance in the global linear model
- Approximation of non-linear dependence of the target variable from the independent variables with piecewise linear local regressions, which are computed for local subsets of data across the SOM
- Automatic optimization of sizes of these local data subsets, with parameterization to avoid overfitting
- Automatic iteration of attribute priorities, to optimally represent non-linear dependences in the SOM
- Visualization of the resulting SOM and associated regression coefficients for the local models, as well as of additional statistical estimates, such as significance, explained variance, and residuals
- Charts with mean descriptives of local models, as well as variance estimates and gain factors for the overall model
- Non-linearity diagnostics providing estimates for the overall model, such as error reduction due to the resolution of non-linearities and the “non-linear influences” of independent variables on the target variable

Graphical validation and comparison of models

- Visual and quantitative validation using score charts, gains charts, and prediction error
- Subdivision of the charts into up to 1000 groups
- Comparison of predicted values with actual values for each model
- Display of all models generated from the same data in the same chart
- Performance comparison of different models, using the model data or the test data sets

Viscovery® SOMine 7 – Data Sheet

Definition and application of scoring models

- Guidance for definition and application of scoring models through the dedicated **Score** workflow
- Interactive definition of "score groups" in the model chart for assigning different measures to subsets of data
- Definition of objective functions using the built-in formula language (e.g., by computing a "customer value" from the prediction value, to select the optimal campaign size)
- Application of scoring models to new data marts by batch computation of prediction values
- Automatic random generation and administration of control groups
- Export of prediction results, score group and, optionally, objective function computation to tab-delimited flat-text files (*.txt) and Microsoft Excel files (*.xlsx)

Evaluation of application results

- Specification of additional attributes in the evaluation data mart, containing the actual value (for which the prediction was computed) and, optionally, control group information, for test or evaluation purposes
- Score, gains, and scenario charts for comparison of results from the application with the control group
- Comparison of actual values and predicted values for the application, control group, or entire data
- Simultaneous display in the charts of further attributes, which are available in the evaluation data mart

Viscovery® SOMine 7 – Data Sheet

Extension Module | Enterprise Data

Connectivity with enterprise environments and support of voluminous data

Main Functions and Features

General characteristics

- **Enterprise Data** is an extension module of the Viscovery SOMine Suite. It requires the **Visual Clusters** core module and can optionally be combined with other modules of the suite.
- This module provides features to connect with enterprise data sources, to handle very high-dimensional data and to support the work process in complex analysis projects.
- The module enables processing of data sets with an unlimited number of records and variables in the complete suite.

Database connectivity

- ODBC/OLEDB interface for accessing all common database systems, such as Oracle and SQL Server
- Import data in the “Import Data” step from database tables and views
- Export of classification and scoring results and of Viscovery data marts to a database table

Additional data interfaces

- Import and export of SPSS files (*.sav) and of Viscovery XML files (*.xml)
- Import data from comma-separated text files (*.csv)
- Import of original SOM_PAK model files (*.cod)

Export of Viscovery data marts

- Export of data marts created in the **Preprocess** workflow to flat files or database tables
- Optional use of defined replacements and transformations for exported attributes
- Optional replacement of missing values by self-organizing map (SOM) node values for exported attributes
- Optional export of nominal variables as dichotomous values

Join functionality

- Joining of data files and/or database tables directly in the “Import Data” step
- Join of data from multiple data sources on arbitrary columns (“left outer join”)
- Highlighting of possible join attributes and name conflicts

Handling of voluminous and high-dimensional data

- Random sampling of large data sets to smaller sizes in Viscovery data marts
- Application of Benjamini-Hochberg multiple-testing correction to adjust statistical confidence measures if a large number of attributes are used to characterize clusters or to compare groups

Preprocessing protocol

- Import and export of preprocessing settings of the **Preprocess** workflow from/to a spreadsheet document
- Compact overview of all parameters (“preprocessing protocol”), which can be edited outside the Viscovery system
- Organization into a table with columns of predefined content, which allows specification of variable definitions, setting allowed variable ranges, defining the treatment of outliers and irregularities, and addition of descriptions, supporting a clear specification of preprocessing options

Viscovery® SOMine 7 – Data Sheet

Viscovery SOMine | Workflow Automation Services

Automated update and application of predictive models at specified times

Main Functions and Features

General characteristics

- The **Workflow Automation Services** constitute a tool package available to complement Viscovery SOMine modules. The package provides dedicated services for the automated execution of each of the five workflow types of Viscovery SOMine.
- User interaction with the tool package is performed through the Viscovery SOMine interface. Scheduled services run on behalf of the user, but do not require the user to be logged in on the computer.

Available services

- The **Preprocess Workflow Automation Service** is based on the **Visual Clusters** core module and automates the creation of a Viscovery data mart from specified input data, using the defined preprocessing settings.
- The **Cluster Workflow Automation Service** is based on the **Visual Clusters** core module and automates the creation of a Viscovery SOM model from a specified data mart, using the defined workflow settings.
- The **Classify Workflow Automation Service** is based on the **Explore and Classify** extension module and automates the application of a segmentation model to a specified application data mart.
- The **Predict Workflow Automation Service** is based on the **Predict and Score** extension module and automates the creation of a prediction model from a specified data mart, using the workflow settings.
- The **Score Workflow Automation Service** is based on the **Predict and Score** extension module and automates the application of a scoring model to an application data mart, according to workflow settings.

Automated execution of workflows

- Automated model creation and application “in the background” at scheduled times
- Execution of workflows as “tasks” on behalf of the user without need for further supervision
- Update of data marts and predictive models using new data from defined data sources
- Application of classification or scoring models to relevant data and export of results to defined targets
- Consistent synchronization of interdependent tasks during automatic execution

Scheduling of tasks

- Task creation by dragging a completed workflow from the project pane to the workflow automation pane
- Specification of input data source and output data target for each task
- Single or periodic execution at scheduled times (monthly, weekly, daily, arbitrarily periodic, or once only)
- Specification of the time of first and last execution
- Definition of possible dependences among listed tasks (result of one task serves as input for other tasks)
- “Start now” and “Cancel task” options for starting and stopping task processing immediately
- Re-use or modification of completed tasks by dragging to the project pane as workflows
- Specification of email address for sending notifications

Presentation and reporting

- Calendar display for daily, monthly or daily overview of scheduled tasks
- Display of dependences among tasks
- Highlighting of task status (scheduled, currently active, completed, disabled, canceled, or failed)
- Retrieval of reports of all workflow steps in completed tasks
- Notification by email after task execution, including result status and report