Data/text mining techniques in modelling the climate effects on crops

Subana Shanmuganathan PhD
Geoinformatics Research Centre (GRC)
School of Computer and Mathematical Sciences (SCMS)
Auckland University of Technology (AUT), New Zealand
December 2013
from data mining and knowledge discovery (in databases) to big data analytics and knowledge extraction; for applications in science

GRC research projects
overview

background
(what’s DM, KD (KDD), big data, data analytics and data science)
• constraints/ challenges
• initiatives: techniques, algorithms needed
• “hot” topics

where are we today?

big data analytics / data science applications
• in enterprises
• disciplines/ problem domains
• new approaches

examples of DM (GRC research)
• climate change effects on grapevine phenology and wine quality
• multi-sensor data analysis
• spatial data mining
Is this the place to learn about mining?

http://cadeh.com/

“Does this count as big data?”

http://magnus-notitia.blogspot.co.nz/2013/02/big-data-is-dead-whats-next.html
"Google was able to spot trends in the Swine Flu epidemic roughly two weeks before the Center for Disease Control by analyzing searches that people were making in different regions of the country."
http://radar.oreilly.com/2010/06/what‐is‐data‐science.html
What’s DM & KD(D)

DM : Data mining is a step in the KDD process consisting of applying data analysis and discovery algorithms that, under acceptable computational efficiency limitations, produce a particular enumeration of patterns over the data (Fayyad et al., 1996).

KD(D) : Knowledge Discovery in Databases is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data (Fayyad et al., 1996).

Big data: Big data is a buzzword, or catch-phrase, used to describe massive volume of both structured and unstructured data that is so large that it's difficult to process using traditional database and software techniques (www.webopedia.com/).

Big data refers to large, diverse, complex, longitudinal, and/or distributed data sets generated from instruments, sensors, Internet transactions, email, video, click streams, and/or all other digital sources available today and in the future.
Big data analytics and Data science

Big data analytics: Big data analytics is the process of examining large amounts of data of a variety of types (big data: 3 Vs: volume, variety, velocity, and lately added veracity) to uncover hidden patterns, unknown correlations and other useful information (actionable knowledge). Such information can provide competitive advantages over rival organisations and result in business benefits, such as more effective marketing and increased revenue.

http://searchbusinessanalytics.techtarget.com/definition/big-data-analytics

Data science: Data Science deals with the whole process of gathering data, pre-processing them and finally making sense out of them, producing what can be called as data products. Large volumes of noisy and unstructured data generated in our daily lives, from social media to search terms on Google cannot be analysed using traditional data mining and warehousing strategies with such large and dynamic data sources.

The need for far more advanced ....for scientific research is historically significant.
Recent NSF **big data** initiatives

Core Techniques and Technologies or fundamental advances in following relevant disciplines:

- computer science
- computational science
- statistics and
- mathematics

deals with research challenges relating to 3 themes:

- data collection and management (DCM)
- analytics
- collaborative environments

NSF: National Science Foundation, USA

http://www.nsf.gov/events/event_summ.jsp?cntn_id=124058&org=CISE
http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504767
Data collection & management

DCM relates to

- types of research being conducted in data management, information retrieval at an expanded scale

need new

- techniques for storing data, new ways of doing I/O
- news architectures to deal with heterogeneous data types
- methods for archiving, indexing and recovering data
- dealing with streaming data
- dealing data already in some complex form
- using various techniques for retrieving the amount of bandwidth for retrieving information from remote data sources
- various techniques for automated annotation
- discovering data sources
- Languages and tools for programming data related obligations
Data analytics

focusses on

– scalable data mining (DM), machine learning (ML) algorithms, statistical inference techniques for dealing with extremely massive and high dimensional, highly heterogeneous and dynamic data sets
– new techniques for predictive modelling such data
– algorithms, programming languages and data structures to deal...
– data-driven simulations techniques for mixing data with simulation, simulation with formal models, information extraction from unstructured and multi modal types of data
– scalable techniques for data visualisation in real time
– techniques for dealing with memory problems that constraints with having to do with big data analytics
Collaborative environments

add science to this:

– data analytics and interpretation become highly interdisciplinary requiring collaborations that

– need techniques for representations, new modelling techniques, and tools that allow for collaborations across individuals looking at complex data sets or across disciplines using multiple representations that make sense within the respective disciplines

– these are Foundational aspects of an effective cyber infrastructure and basic problems have to be addressed in the respective disciplines
Topics

Biomedical applications

– Various techniques for analysing structural and functional correlatives, interactions and networks, various protein interaction networks, network of neurons
– Example analysis on social media for understanding local, original and national health
– New techniques for mining literature and other types of data to get an understanding of the biomedical research landscape, techniques for analysing multiple clinical research data sets
– Predictive modelling in biology are related to human health and treating disease
– Clinical science to generate hypotheses using already available background knowledge
– New techniques for disseminating scientific knowledge beyond traditional publications. Methods to link the publications to data sets, simulations → one can actually replicate the study reported in the publication
Artemis platform

big data and data mining

Data Mining

Big Data

Lots of detail

A close up view

The big picture

Lots of relationships

Peter Cochrane, Cochrane Associates on Jan 31, 2013
http://www.slideshare.net/PeterCochrane/big-data-v-data-mining
A knowledge-discovery life-cycle for Big Data

Strawman compute-intensive vs. data-intensive computer architectures in the 2017 timeframe

A comparison of selected performance parameters for different benchmarks with data analytics and mining workloads

Generalized Scientific Workflow

<table>
<thead>
<tr>
<th>Data Generation Phase (scenarios)</th>
<th>Overview</th>
<th>Transactional / in situ post processing requirements</th>
<th>Storage for Post processing</th>
<th>Sharing and distribution</th>
<th>Visualization</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) Design Oriented Exascale Simulations e.g., combustion, CFD</td>
<td>Generation of data from simulations</td>
<td>Data reduction for post processing (1) feature detection &amp; tracking (3) advanced analytics</td>
<td>Reduced data Low (Only the producer or a few scientists may analyze data in the future)</td>
<td>In-situ, interaction, feature display, uncertainty, visual debugging</td>
<td></td>
</tr>
<tr>
<td>2) Discovery Oriented Exascale Simulations (2) e.g., climate, cosmology</td>
<td>Integration of data generated from simulation and observations</td>
<td>Data reduction for post processing (1) raw data (2) Wall organized (DB) (3) enriched for queries</td>
<td>High (A large number of scientists, geographically distributed)</td>
<td>InfoVis and SciVis, pattern detection, correlation, clustering, ensemble vis, uncertainty</td>
<td></td>
</tr>
<tr>
<td>3) Large centralized instruments e.g., LHC</td>
<td>Data generation from large devices Extremely high rates Centralized, coordinated, controlled access</td>
<td>HW/SW for high-rate data processing (1) Derived data (2) metadata (3) Extensive queries</td>
<td>Raw data (1) Different forms of derived data (5) Lots of distributed copies</td>
<td>High (A large number of scientists, geographically distributed), different sets defined by queries and other parameters</td>
<td>Custom user interfaces enabling query visual analysis, trajectory vis/analysis, user driven data triage / summarization</td>
</tr>
<tr>
<td>4) Smaller distributed instruments e.g., field work, sensors, biology</td>
<td>Data generation from massive numbers of distributed devices, sensors</td>
<td>Local processing and derivations (1) Integration of massive data (possibly at an exascale level system, data centers)</td>
<td>Raw data (1) Derived data and subsets (3) Distributed copies</td>
<td>High (A large number of scientists, geographically distributed)</td>
<td>InfoVis, high dimensional vis, large-scale graphs, patterns, clustering, scalability</td>
</tr>
</tbody>
</table>

Table 3.1: Data generation requirements for different domains
Figure 4.3: The timing breakdown (in seconds) for in-situ, in-transit, and data movement for the simulation and the various analytics algorithms using 4896 cores on Jaguar, the Cray XK6 at Oak Ridge National Laboratory’s National Center for Computational Sciences. 4480 cores were used for the simulation and in-situ processing, 256 cores were used for in-transit processing and 160 cores were used for task scheduling and data movement. The simulation grid size was 1600x1372x430 and all measurements are per simulation time step.
where are we today?

preparing for post Moore Era: race for next level performance/Exascale systems

– need it by the end of this decade/early next
– to achieve progress in the simulation of
  • societal impacts of weather, environmental change
  • continued certification of nuclear stockpile
  • combustion simulation
  • national security
  • to begin to understand brain functioning

DRAPA launched UHPC research aimed @ achieving petascale performance in a single rack system consuming only 57KW

DRAPA: Defence Advanced Research Projects Agency
UHPC: Ubiquitous High Performance Computing
Exascale Research: Preparing for the Post Moore Era Marc Snir, William Gropp and Peter Kogge 6/19/11
https://www.ideals.illinois.edu/bitstream/handle/2142/25469/Exascale%20Research.pdf?sequence=2
Achieving Exascale X1000 in 2015?

Not just high end, floating point intensive, supercomputers (“exoflops” machines) but across the board 3 classes 1) data center-sized systems 2) departmental-sized systems and 3) embedded systems

**Identified challenges**

- Energy and Power Challenge
- Memory and Storage Challenge
- Concurrency and Locality Challenge
- Resiliency Challenge

**More research reqd in Co-development and optimization of Exascale**

- Hardware Technologies and Architectures
- Architectures and Programming Models
- Algorithms, Applications, Tools, and Run-times
- Development of a deep understanding of how to architect Resilient Exascale Systems

**Suggested 3 phased research agenda**

- A System Architecture Exploration Phase
- A Technology Demonstration Phase
- A Scalability Slice Prototyping Phase

ExaScale Computing Study: Technology Challenges in Achieving Exascale Systems

current CMOS won’t last

CMOS technology is slowing down

– Stein’s Law... something cannot go on forever, forecasts a feature size of 7.5 nm by 2024

– will plateau next decade and no alternative is ready yet i.e., spintronics, Rapid Single Flux Quantum (RSFQ) Logic (requires cryogenic cooling)
Why we need exascale

http://www.bbc.co.uk/news/health-24428162
Synergistic Challenges in Data-Intensive Science and Exascale Computing

Figure 3.2: The surface of the earth is warming. Four maps the trend, with average surface temperature for 1919 (top left) and 2099 (bottom right) shown. The global average, as a line. The computations were based on observed temperature (e.g., multi-model ensemble) assuming a continued:

Figure 4.2: Federated data access in climate workflow
Figure 4.6: The visualization on the left shows features for a single time step within a large-scale combustion simulation while the figure on the right shows a graph that tracks features over time. Combustion scientists can follow specific events over time. Simulation by Jackie Chen, Sandia National Laboratories; Visualization by the Scientific Computing and Imaging (SCI) Institute.
Figure 4.8: The rise of the fields of computational biology and bioinformatics has brought significant advances in algorithms for processing biological datasets. However, deciphering raw data and computational results through visual representations is too often done as the final step in a complex research process, with tools that are rarely specific to the task. Hence, there is a significant opportunity to enhance biological data analysis through a thoughtful and principled investigation of visualization. These four tools are examples of custom, interactive visualization tools designed in collaborations with biologists—they have all been deployed in biological research labs and led to a variety of scientific insights [14].
Data science

Data-Drive Decision Making
(across the firm)

Automated DDD

Data Science

Data Engineering and Processing
[including "Big Data" technologies]

Other positive effects of data processing
(e.g. faster transaction processing)
http://www.youtube.com/watch?v=duC4PDObFWU#t=16
Open Source Apache Hadoop

BIG ANALYTICS

End Users

Apps

BI & Visualization

Business Analytics

Data Analytics Use

SQL, Data Flow Languages,
Predictive Analytics

BIG DATA

Operational Data

Unstructured Data

Structured Data

ETL & Data Integration Products

Workflow / Scheduling Products

System Tools

Hadoop (MapReduce & HDFS)

NoSQL (Hadoop Based)

Hadoop

MapReduce, HDFS, Serialization,
Coordination,
Scalable database ...

Hadoop is an Apache Software Foundation collection of projects
Step 1: Put a bunch of data in Hadoop

BIG ANALYTICS
- Developer Environments
  - Developers
- Analytics Products
  - Data Analysts

End Users
- Apps
- BI & Visualization Tools
- Business Analysis

Data Analytics & Use

BIG DATA
- Unstructured Data
- Structured Data
- Integration Products
  - Workflow / Scheduler Products
  - System Tools

Hadoop
- MapReduce & HDFS
  - NoSQL (Hadoop Based)

Operational Data

Data Management & Storage
Step 2: Do some one-off big picture analytics and use what you learn to inform next steps.

**BIG ANALYTICS**
- Developer Environments
- Analytic Products
- End Users
- Apps
- BI & Visualization Tools
- Business Analysts

**BIG DATA**
- Hadoop (MapReduce & HDFS)
- NoSQL (Hadoop Based)
- Operational Data
- Structured Data
- ETL & Integration Products
- Workflow/Scheduler Tools
- System Admins

**Ad Hoc Analysis**

---

Tuesday 2 December 2013
Step 3 Develop some basic indexes to provide a basic search across all the data, put it in NoSQL to support some customer facing apps.
Step 4 Precompute answers to valuable analytics, host results in NoSql
Schema-less format requires very little pre-processing

"Heisenberg principal" applied to data
Traditional Vs modern day fraud detection

Figure 2-1  Traditional fraud detection patterns use approximately 20 percent of available data.

Figure 2-2  A modern-day fraud detection ecosystem synergizes a Big Data platform with traditional processes.

Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data: page 22-23
Tuesday 2 December 2013
<table>
<thead>
<tr>
<th>Evolutionary Step</th>
<th>Business Question</th>
<th>Enabling Technologies</th>
<th>Product Providers</th>
<th>Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Collection (1960s)</td>
<td>&quot;What was my total revenue in the last five years?&quot;</td>
<td>Computers, tapes, disks</td>
<td>IBM, CDC</td>
<td>Retrospective, static data delivery</td>
</tr>
<tr>
<td>Data Access (1980s)</td>
<td>&quot;What were unit sales in New England last March?&quot;</td>
<td>Relational databases (RDBMS), Structured Query Language (SQL), ODBC</td>
<td>Oracle, Sybase, Informix, IBM, Microsoft</td>
<td>Retrospective, dynamic data delivery at record level</td>
</tr>
<tr>
<td>Data Warehousing &amp; Decision Support (1990s)</td>
<td>&quot;What were unit sales in New England last March? Drill down to Boston.&quot;</td>
<td>On-line analytic processing (OLAP), multidimensional databases, data warehouses</td>
<td>Pilot, Comshare, Arbor, Cognos, Microstrategy</td>
<td>Retrospective, dynamic data delivery at multiple levels</td>
</tr>
<tr>
<td>Data Mining (Emerging Today)</td>
<td>&quot;What’s likely to happen to Boston unit sales next month? Why?&quot;</td>
<td>Advanced algorithms, multiprocessor computers, massive databases</td>
<td>Pilot, Lockheed, IBM, SGI, numerous startups (nascent industry)</td>
<td>Prospective, proactive information delivery</td>
</tr>
</tbody>
</table>

GRC research projects

1. climate change effects on grapevine phenology and wine quality
2. multi-sensor data analysis
3. Pixel clustering for spatial data mining
Climate change effects on grapevine phenology and wine quality

Precise and structured

Video, text, ratings
Audio, web

Precise and structured

1000 years old diary

sommelier comments...
Issues and solutions

Extract data from web portal

Data/text pre-processing

data Dimension reduction

Analysis

Tuesday 2 December 2013
Text mining-vector space model

Remove stop words

Stemmer

Create matrix

Feature/ descriptor extraction

K-means

SOM

Feature selection: study the clusters and their profiles

PCA

\[ \hat{x}_{k-1} = x - \sum_{i=1}^{k-1} w_i w_i^T x \]

\[ w_i = tf_i \times \log \left( \frac{D}{df_i} \right) \]

\[ F = \sum_{v=1}^{k-1} tf \times idf \]

\[
\begin{array}{cccccccc}
Doc & aa & ab & \ldots & \ldots & \ldots & zoo \\
1 & 0 & 2 & 0 & 0 & 0 & 0 & 0 \\
2 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\
3 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
\end{array}
\]
Text mining: Sommelier comments

WineEnthusiast.com

**WINE ENTHUSIAST MAGAZINE**

*Viu Manent 2007 Reserva Chardonnay (Casablanca Valley)*
For a first effort from Casablanca, Viu Manent has hit a home run. This wine is a classic New World Chard, meaning it’s liberally oaked, vibrant, ripe and full of tropical fruit. But along with the obvious there are also notes of cinnamon, mineral, exotic apple and butterscotch. Imported by Baystate Wine Co. — M.S. Published 7/1/2008

Best Buy!

90 points

WineEnthusiast.com

**WINE ENTHUSIAST MAGAZINE**

*Undurraga 2005 Aliwen Reserva Chardonnay (Central Valley)*
This new wine from a venerable Chilean producer scores points all over the map. The nose is a smooth ride of white fruits and cleanliness, while the mouth pleases with pear, green

90 points
Web text mining wine comments
The WEBSOM approach to Marlborough wine styles
WEBSOM of 195 descriptors extracted from sommelier comments provided for 253 Marlborough vintages of styles produced from this famous wine region. The different segments in the SOM show the descriptors used to state the features (S1-S22) of the vintages by sommeliers.


S19: herbal-88 tropic-182
butter-29 delic-53 fine-64 floral-69 linger-101 oak-119 smoke-155 smoki-156 spice-161
spici-162 subtl-167 toast-180 vanilla-183 - **Chardonnay**
derri-20 caramel-32 chocol-37 cinnamon-38 clove-42 coffe-43 complex-45 dark-52 dri-57
firm-65 mushroom-115 plum-134 readi-141 roast-147 silki-152 smooth-157 structur-166
suppl-169 tannin-174 velveti-186 vintag-187 – **Pinot Noir**
chalki-34 miner-111 nectarin-116 pink-132 pungent-136 sweati-170 white-190 winemak-191
dry-58 honei-89 riesl-145
bottl-26 creami-49 rich-144
acid-1 appl-8 balanc-17
gooseberri-74 lean-95 lime-100 raci-139 tart-175
light-99 modest-113 simpl-153 solid-159
bodi-25 full-73 medium-108
green-78 pepper-127 refresh-143
pineappl-131
citru-40 peach-125 ripe-146
intens-91
pear-126
sweet-171
bright-27 fruiti-72 soft-158
herbal-88 tropic-182
clean-41 crisp-50 fresh-71 grapefruit-75 herb-86 melon-109
black-22 cherri-36 cola-44 noir-117 pinot-133
grassi-77
NZ Chardonnay

Gisborne
sweet-19 <= 0
  |  spice-18 <= 0
  |  |  appl-1 <= 0.27: med
    (28.0/7.0)
  |  |  appl-1 > 0.27: high (2.0)
  |  spice-18 > 0: high (3.0/2.0)
sweet-19 > 0
  |  vanilla-23 <= 0: med (3.0)
  |  vanilla-23 > 0: low (3.0)

Hawke’s Bay
lime-19 <= 0
  |  ripe-28 <= 0.23
  |  |  orang-23 <= 0
  |  |  |  creami-9 <= 0
  |  |  |  |  honei-17 <= 0
  |  |  |  |  |  intens-18 <= 0: med
    (19.0/3.0)
  |  |  |  |  intens-18 > 0: high (5.0/1.0)
  |  |  |  |  honei-17 > 0: high (2.0)
  |  |  |  |  creami-9 > 0: high (2.0)
  |  |  |  orang-23 > 0: high (3.0)
  |  ripe-28 > 0.23: med (8.0/1.0)
lime-19 > 0: med (6.0/1.0)

Training: 76%
Cross validation : 46%

Waipara
toast-8 <= 0.26
  |  citru-3 <= 0: med (8.0/2.0)
  |  citru-3 > 0: high (2.0/1.0)
toast-8 > 0.26: high (3.0)

Training: 76%
Cross validation : 38%

low <80 medium (med) >79 and <90 high >89 (100 point)
## Chardonnay

<table>
<thead>
<tr>
<th>region</th>
<th>low</th>
<th>med</th>
<th>high</th>
</tr>
</thead>
<tbody>
<tr>
<td>Waipara</td>
<td></td>
<td></td>
<td>toast, citrus</td>
</tr>
<tr>
<td>Gisborne</td>
<td>vanilla</td>
<td></td>
<td>apple, spice</td>
</tr>
<tr>
<td>Hawke’s Bay</td>
<td></td>
<td></td>
<td>Intense, honey,</td>
</tr>
</tbody>
</table>
<pre><code>                      |        |        | creamy, orange    |
</code></pre>
Climate Vs Viticulture

Base climate
• 3-5 decade average
• Used for Grapevine variety selection

Annual (year to year)
• Determines vintage quality
• Responsible for phenology

precise data Vs subjective wine quality
Spatiotemporal scales & climate change modelling

**Fig. 1 Climatic scales related to surface and time.**

Kumeu River Wines, New Zealand

Based on annual yield high (1) / med (2) / low (3) yield years are determined
Daily extreme weather matrix

Temperature intervals

<table>
<thead>
<tr>
<th></th>
<th>&lt;8.1-11.00</th>
<th>11.1-13.00</th>
<th>13.1-16.00</th>
<th>16.1-19.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>W1</td>
<td>4 (frequency)</td>
<td>3</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>W2</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>

Tuesday 2 December 2013
$\chi^2$ test: Daily max temp freq matrix
\( \chi^2 \) test on daily ext max T

<table>
<thead>
<tr>
<th>week</th>
<th>(&lt;23)</th>
<th>23.1-26</th>
<th>(&gt;26)</th>
<th>rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>31</td>
<td>17.33</td>
<td>2.67</td>
<td>1.00</td>
<td>1 low yield</td>
</tr>
<tr>
<td>31</td>
<td>11.67</td>
<td>8.00</td>
<td>1.33</td>
<td>3 high yield</td>
</tr>
<tr>
<td>32</td>
<td>17.67</td>
<td>3.00</td>
<td>0.33</td>
<td>1</td>
</tr>
<tr>
<td>32</td>
<td>8.67</td>
<td>9.00</td>
<td>3.33</td>
<td>3</td>
</tr>
<tr>
<td>33</td>
<td>16.00</td>
<td>4.67</td>
<td>0.33</td>
<td>1</td>
</tr>
<tr>
<td>33</td>
<td>8.33</td>
<td>8.33</td>
<td>4.33</td>
<td>3</td>
</tr>
<tr>
<td>34</td>
<td>11.33</td>
<td>7.00</td>
<td>2.67</td>
<td>1</td>
</tr>
<tr>
<td>34</td>
<td>7.67</td>
<td>7.67</td>
<td>5.67</td>
<td>3</td>
</tr>
<tr>
<td>35</td>
<td>6.00</td>
<td>9.33</td>
<td>5.67</td>
<td>1</td>
</tr>
<tr>
<td>35</td>
<td>5.33</td>
<td>9.33</td>
<td>6.33</td>
<td>3</td>
</tr>
<tr>
<td>36</td>
<td>4.00</td>
<td>9.67</td>
<td>7.33</td>
<td>1</td>
</tr>
<tr>
<td>36</td>
<td>4.67</td>
<td>8.67</td>
<td>7.67</td>
<td>3</td>
</tr>
<tr>
<td>37</td>
<td>5.00</td>
<td>8.00</td>
<td>8.00</td>
<td>1</td>
</tr>
<tr>
<td>37</td>
<td>3.67</td>
<td>8.67</td>
<td>8.67</td>
<td>3</td>
</tr>
<tr>
<td>38</td>
<td>5.00</td>
<td>6.00</td>
<td>10.00</td>
<td>1</td>
</tr>
<tr>
<td>38</td>
<td>5.33</td>
<td>7.00</td>
<td>8.67</td>
<td>3</td>
</tr>
<tr>
<td>38</td>
<td>5.00</td>
<td>5.33</td>
<td>10.67</td>
<td>1</td>
</tr>
</tbody>
</table>
\( \chi^2 \) test results daily ext max T

<table>
<thead>
<tr>
<th>week No.</th>
<th>&lt;23</th>
<th>23.1-26</th>
<th>&gt;26</th>
<th>chi square rate</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>31</td>
<td>11.67</td>
<td>8.00</td>
<td>1.33</td>
<td>8.000</td>
<td>0.005</td>
</tr>
<tr>
<td>32</td>
<td>17.67</td>
<td>3.00</td>
<td>0.33</td>
<td>9.228</td>
<td>0.002</td>
</tr>
<tr>
<td>32</td>
<td>8.67</td>
<td>9.00</td>
<td>3.33</td>
<td>7.364</td>
<td>0.007</td>
</tr>
<tr>
<td>33</td>
<td>16.00</td>
<td>4.67</td>
<td>0.33</td>
<td>7.247</td>
<td>0.007</td>
</tr>
<tr>
<td>33</td>
<td>8.33</td>
<td>8.33</td>
<td>4.33</td>
<td>10.286</td>
<td>0.001</td>
</tr>
<tr>
<td>44</td>
<td>5.33</td>
<td>8.00</td>
<td>7.67</td>
<td>6.125</td>
<td>0.013</td>
</tr>
<tr>
<td>45</td>
<td>4.33</td>
<td>7.33</td>
<td>9.33</td>
<td>14.235</td>
<td>0.000</td>
</tr>
<tr>
<td>45</td>
<td>9.00</td>
<td>10.00</td>
<td>2.00</td>
<td>4.900</td>
<td>0.027</td>
</tr>
</tbody>
</table>

Low yield
High yield
## Pinot Gris

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>&lt;23°C</th>
<th>23-26 °C</th>
<th>&gt;26 °C</th>
<th>X²</th>
<th>p-value</th>
<th>h/l</th>
</tr>
</thead>
<tbody>
<tr>
<td>13N-4De</td>
<td>30</td>
<td>13</td>
<td>7.5</td>
<td>0.5</td>
<td>3.86</td>
<td>0.050</td>
<td>low</td>
</tr>
<tr>
<td>4-24Dec</td>
<td>33</td>
<td>12</td>
<td>7.5</td>
<td>0.5</td>
<td>6.82</td>
<td>0.009</td>
<td>low</td>
</tr>
<tr>
<td>1-20Jan</td>
<td>37</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>4.00</td>
<td>0.046</td>
<td>low</td>
</tr>
<tr>
<td>12-22Feb</td>
<td>43</td>
<td>6</td>
<td>7.5</td>
<td>7.5</td>
<td>4.00</td>
<td>0.046</td>
<td>high</td>
</tr>
<tr>
<td>19F-11M</td>
<td>44</td>
<td>9.5</td>
<td>7.5</td>
<td>4</td>
<td>5.54</td>
<td>0.190</td>
<td>high</td>
</tr>
<tr>
<td>3.5</td>
<td>7</td>
<td>10.5</td>
<td>5.83</td>
<td>0.016</td>
<td>low</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26F-18M</td>
<td>45</td>
<td>11.5</td>
<td>8</td>
<td>1.5</td>
<td>4.24</td>
<td>0.400</td>
<td>high</td>
</tr>
<tr>
<td>9.5</td>
<td>11.64</td>
<td>0.001</td>
<td>low</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Early Feb mid March (berry ripening) → < 23 °C produces high yield and >26 °C leads to low yield.
<table>
<thead>
<tr>
<th>Char</th>
<th>wk</th>
<th>&lt;23 °C</th>
<th>23-26 °C</th>
<th>&gt;26 °C</th>
<th>X²</th>
<th>p-value</th>
<th>h/l</th>
</tr>
</thead>
<tbody>
<tr>
<td>11-31No</td>
<td>31</td>
<td>18.0</td>
<td>13.0</td>
<td>0.5</td>
<td>7.11</td>
<td>0.008</td>
<td>high</td>
</tr>
<tr>
<td>18N-7De</td>
<td>32</td>
<td>15.0</td>
<td>13.0</td>
<td>3.5</td>
<td>6.08</td>
<td>0.014</td>
<td>high</td>
</tr>
<tr>
<td>2-22Dec</td>
<td>34</td>
<td>16.0</td>
<td>10.5</td>
<td>5.0</td>
<td>4.59</td>
<td>0.032</td>
<td>high</td>
</tr>
<tr>
<td>9-29Dec</td>
<td>35</td>
<td>12.0</td>
<td>15.5</td>
<td>4.0</td>
<td>5.77</td>
<td>0.016</td>
<td>high</td>
</tr>
<tr>
<td>16D-5Ja</td>
<td>36</td>
<td>9.0</td>
<td>15.5</td>
<td>7.0</td>
<td>7.35</td>
<td>0.007</td>
<td>high</td>
</tr>
<tr>
<td>20J-9Feb</td>
<td>41</td>
<td>4.5</td>
<td>6.5</td>
<td>9.0</td>
<td>6.40</td>
<td>0.011</td>
<td>low</td>
</tr>
<tr>
<td>27J-6Feb</td>
<td>42</td>
<td>4.5</td>
<td>9.0</td>
<td>6.5</td>
<td>6.40</td>
<td>0.011</td>
<td>low</td>
</tr>
<tr>
<td>3-23 Feb</td>
<td>43</td>
<td>2.5</td>
<td>17.5</td>
<td>11.5</td>
<td>5.45</td>
<td>0.020</td>
<td>high</td>
</tr>
<tr>
<td>17F-9 Ma</td>
<td>45</td>
<td>3.0</td>
<td>6.5</td>
<td>13.5</td>
<td>4.12</td>
<td>0.042</td>
<td>low</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9.0</td>
<td>15.5</td>
<td>7.0</td>
<td>6.0</td>
<td>0.014</td>
<td>high</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7.35</td>
<td></td>
<td></td>
<td></td>
<td>0.007</td>
<td>high</td>
</tr>
</tbody>
</table>

early Feb early March (berry ripening) $\rightarrow$ < 26°C produces high yield
## Pinot Noir

<table>
<thead>
<tr>
<th>PN</th>
<th>wk</th>
<th>&lt;23 °C</th>
<th>23-26 °C</th>
<th>&gt;26 °C</th>
<th>( X^2 )</th>
<th>p-value</th>
<th>h/l</th>
</tr>
</thead>
<tbody>
<tr>
<td>12N-1Dec</td>
<td>32</td>
<td>17</td>
<td>3.5</td>
<td>0.5</td>
<td>4.50</td>
<td>0.034</td>
<td>high</td>
</tr>
<tr>
<td>19N-8Dec</td>
<td>33</td>
<td>13</td>
<td>7.5</td>
<td>0.5</td>
<td>9.94</td>
<td>0.002</td>
<td>high</td>
</tr>
<tr>
<td>26N-15De</td>
<td>34</td>
<td>17.5</td>
<td>2.0</td>
<td>1.5</td>
<td>4.74</td>
<td>0.029</td>
<td>low</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9.5</td>
<td>8.5</td>
<td>3.0</td>
<td>8.05</td>
<td>0.005</td>
<td>high</td>
</tr>
<tr>
<td>3-23 Dec</td>
<td>35</td>
<td>16.0</td>
<td>4.0</td>
<td>1.0</td>
<td>12.90</td>
<td>0.000</td>
<td>low</td>
</tr>
<tr>
<td></td>
<td>35</td>
<td>4.5</td>
<td>12.0</td>
<td>4.5</td>
<td>8.00</td>
<td>0.005</td>
<td>high</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4.46</td>
<td>0.035</td>
<td>high</td>
</tr>
<tr>
<td>17D-6Jan</td>
<td>37</td>
<td>11.5</td>
<td>8.0</td>
<td>1.5</td>
<td>4.24</td>
<td>0.400</td>
<td>low</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5.5</td>
<td>9.5</td>
<td>6.0</td>
<td>5.40</td>
<td>0.020</td>
<td>high</td>
</tr>
<tr>
<td>24D-13Ja</td>
<td>38</td>
<td>5.0</td>
<td>8.5</td>
<td>7.5</td>
<td>6.37</td>
<td>0.012</td>
<td>high</td>
</tr>
<tr>
<td>31D-20Ja</td>
<td>39</td>
<td>7</td>
<td>9</td>
<td>5</td>
<td>5.56</td>
<td>0.018</td>
<td>low</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>8</td>
<td>11</td>
<td></td>
<td>4.50</td>
<td>0.034</td>
<td>high</td>
</tr>
<tr>
<td>4-24Feb</td>
<td>44</td>
<td>4</td>
<td>6.5</td>
<td>10.5</td>
<td>5.44</td>
<td>0.020</td>
<td>high</td>
</tr>
<tr>
<td>11F-3Mar</td>
<td>45</td>
<td>6</td>
<td>7.5</td>
<td>7.5</td>
<td>9.32</td>
<td>0.002</td>
<td>high</td>
</tr>
</tbody>
</table>

early Feb early March (berry ripening) → < 23 °C produces high yield
$\chi^2$ test results/ graphs daily ext max T

Flowering (pollination)

Berry ripening

Tuesday 2 December 2013
MULTI SENSOR DATA
live web display
The methodology

The multi-sensor data

1. id
2. DateTime
3. NodeId
4. Pressure_Rel
5. Ind_Temp
6. Ind_Hum
7. Out_Temp
8. Out_Hum
9. Dewp
10. Windc
11. Winds
12. Wind_Dir
13. Gust
14. Rain_Rate
15. Act_rain
16. Rain_Today
17. Pressure_Abs
18. VinyardId
19. Rain_Total
20. Heat_Indx
21. High_Gust

Record high temperature 32.8°C
Record low temperature -8.9°C
Record high gust 172.2 km/h
Record high average 172.2 km/h
Record daily rain 82.6 mm
Record low wind chill -11.5°C
Record high barometer 1035.6 hPa
Record low barometer 977.9 hPa

From
http://www.binoscope.co.nz/Kumeu.htm

1. id
2. Date Time
3. Pressure_Rel
4. Out_Temp
5. Out_Hum
6. Dewp
7. Windc
8. Winds
9. Wind_Dir
10. Gust
11. class
12. Heat_Indx
13. E_code

Tuesday 2 December 2013
Data distribution

Data mining

• C5.0
• C&RT (classification and regression trees-B)
• CAHID (Chi-squared Automatic Interaction) Detector
• ANN
• Regression
• PCA

Sw: SPSS clementine
C5.0 for high gust
C5.0 for very high gust

```
Rule 1 for ve (12. 0.75) if pressure_rel > 909 and out_temp > -9.8 and winds <= 4.9 and winds > 1 and heat_index <= 0 and out_tmp
Rule 2 for ve (3. 1.0) if pressure_rel > 909 and out_temp > -9.8 and winds <= 4.9 and winds > 1 and heat_index <= 0 and out_tmp
Rule 3 for ve (2. 1.0) if pressure_rel > 909 and out_temp > -9.8 and winds <= 4.9 and winds > 1 and heat_index <= 0 and out_tmp
Rule 4 for ve (2. 1.0) if pressure_rel > 909 and out_temp > -9.8 and winds <= 4.9 and winds > 1 and heat_index <= 0 and out_tmp
Rule 5 for ve (2. 1.0) if pressure_rel > 909 and out_temp > -9.8 and winds <= 4.9 and winds > 1 and heat_index <= 0 and out_tmp
Rule 6 for ve (2. 1.0) if pressure_rel > 909 and out_temp > -9.8 and winds <= 4.9 and winds > 1 and heat_index <= 0 and out_tmp
Rule 7 for ve (5. 0.6) if pressure_rel > 909 and out_temp > -9.8 and winds <= 4.9 and winds > 1 and heat_index <= 0 and out_tmp
Rule 8 for ve (7. 0.85) if pressure_rel > 909 and out_temp > -9.8 and winds <= 4.9 and winds > 1 and heat_index <= 0 and out_tmp
Rule 9 for ve (4. 1.0) if pressure_rel > 909 and out_temp > -9.8 and winds <= 4.9 and winds > 1 and heat_index <= 0 and out_tmp
Rule 10 for ve (2. 1.0) if pressure_rel > 909 and out_temp > -9.8 and winds <= 4.9 and winds > 1 and heat_index <= 0 and out_tmp
Rule 11 for ve (6. 1.0) if pressure_rel > 909 and out_temp > -9.8 and winds <= 4.9 and winds > 1 and heat_index <= 0 and out_tmp
Rule 12 for ve (6. 0.8333) if pressure_rel > 909 and out_temp > -9.8 and winds <= 4.9 and winds > 1 and heat_index <= 0 and out_tmp
Rule 13 for ve (6. 0.8333) if pressure_rel > 909 and out_temp > -9.8 and winds <= 4.9 and winds > 1 and heat_index <= 0 and out_tmp
Rule 14 for ve (6. 0.75) if pressure_rel > 909 and out_temp > -9.8 and winds <= 4.9 and winds > 1 and heat_index <= 0 and out_tmp
Rule 15 for ve (25. 0.36) if pressure_rel > 909 and out_temp > -9.8 and winds <= 4.9 and winds > 1 and heat_index <= 0 and out_tmp
Rule 16 for ve (3. 0.66) if pressure_rel > 909 and out_temp > -9.8 and winds <= 4.9 and winds > 1 and heat_index <= 0 and out_tmp
Rule 17 for ve (7. 0.857) if pressure_rel > 909 and out_temp > -9.8 and winds <= 4.9 and winds > 1 and heat_index <= 0 and out_tmp
Rule 18 for ve (5. 1.0) if pressure_rel > 909 and out_temp > -9.8 and winds <= 4.9 and winds > 1 and heat_index <= 0 and out_tmp
Rule 19 for ve (43. 0.312) if pressure_rel > 909 and out_temp > -9.8 and winds <= 4.9 and winds > 1 and heat_index <= 0 and out_tmp
Rule 20 for ve (5. 1.0) if pressure_rel > 909 and out_temp > -9.8 and winds <= 4.9 and winds > 1 and heat_index <= 0 and out_tmp
Rule 21 for ve (7. 0.857) if pressure_rel > 909 and out_temp > -9.8 and winds <= 4.9 and winds > 1 and heat_index <= 0 and out_tmp
Rule 22 for ve (5. 1.0) if pressure_rel > 909 and out_temp > -9.8 and winds <= 4.9 and winds > 1 and heat_index <= 0 and out_tmp
Rule 23 for ve (8. 0.625) if pressure_rel > 909 and out_temp > -9.8 and winds <= 4.9 and winds > 1 and heat_index <= 0 and out_tmp
Rule 24 for ve (6. 0.75) if pressure_rel > 909 and out_temp > -9.8 and winds <= 4.9 and winds > 1 and heat_index <= 0 and out_tmp
Rule 25 for ve (10. 0.8) if pressure_rel > 909 and out_temp > -9.8 and winds <= 4.9 and winds > 1 and heat_index <= 0 and out_tmp
Rule 26 for ve (5. 0.8) if pressure_rel > 909 and out_temp > -9.8 and winds <= 4.9 and winds > 1 and heat_index <= 0 and out_tmp
Rule 27 for ve (31. 0.568) if pressure_rel > 909 and out_temp > -9.8 and winds <= 4.9 and winds > 1 and heat_index <= 0 and out_tmp
Rule 28 for ve (3. 1.0) if pressure_rel > 909 and out_temp > -9.8 and winds <= 4.9 and winds > 1 and heat_index <= 0 and out_tmp
Rule 29 for ve (7. 0.857) if pressure_rel > 909 and out_temp > -9.8 and winds <= 4.9 and winds > 1 and heat_index <= 0 and out_tmp
Rule 30 for ve (3. 0.667) if pressure_rel > 909 and out_temp > -9.8 and winds <= 4.9 and winds > 1 and heat_index <= 0 and out_tmp
Rule 31 for ve (11. 0.636) if pressure_rel > 909 and out_temp > -9.8 and winds <= 4.9 and winds > 1 and heat_index <= 0 and out_tmp
Rule 32 for ve (9. 0.667) if pressure_rel > 909 and out_temp > -9.8 and winds <= 4.9 and winds > 1 and heat_index <= 0 and out_tmp
Rule 33 for ve (9. 0.778) if pressure_rel > 909 and out_temp > -9.8 and winds <= 4.9 and winds > 1 and heat_index <= 0 and out_tmp
Rule 34 for ve (20. 0.65) if pressure_rel > 909 and out_temp > -9.8 and winds <= 4.9 and winds > 1 and heat_index <= 0 and out_tmp
Rule 35 for ve (2. 1.0) if pressure_rel > 909 and out_temp > -9.8 and winds <= 4.9 and winds > 1 and heat_index <= 0 and out_tmp
Rule 36 for ve (41. 0.854) if pressure_rel > 909 and out_temp > -9.8 and winds <= 4.9 and winds > 1 and heat_index <= 0 and out_tmp
Rule 37 for ve (5. 1.0) if pressure_rel > 909 and out_temp > -9.8 and winds <= 4.9 and winds > 1 and heat_index <= 0 and out_tmp
Rule 38 for ve (13. 0.74) if pressure_rel > 909 and out_temp > -9.8 and winds <= 4.9 and winds > 1 and heat_index <= 0 and out_tmp
Rule 39 for ve (6. 0.75) if pressure_rel > 909 and out_temp > -9.8 and winds <= 4.9 and winds > 1 and heat_index <= 0 and out_tmp
Rule 40 for ve (14. 0.788) if pressure_rel > 909 and out_temp > -9.8 and winds <= 4.9 and winds > 1 and heat_index <= 0 and out_tmp
Rule 41 for ve (11. 0.74) if pressure_rel > 909 and out_temp > -9.8 and winds <= 4.9 and winds > 1 and heat_index <= 0 and out_tmp
Rule 42 for ve (13. 0.74) if pressure_rel > 909 and out_temp > -9.8 and winds <= 4.9 and winds > 1 and heat_index <= 0 and out_tmp
Rule 43 for ve (14. 0.788) if pressure_rel > 909 and out_temp > -9.8 and winds <= 4.9 and winds > 1 and heat_index <= 0 and out_tmp
Rule 44 for ve (2. 1.0) if pressure_rel > 909 and out_temp > -9.8 and winds <= 4.9 and winds > 1 and heat_index <= 0 and out_tmp
Rule 45 for ve (2. 1.0) if pressure_rel > 909 and out_temp > -9.8 and winds <= 4.9 and winds > 1 and heat_index <= 0 and out_tmp
Rule 46 for ve (2. 1.0) if pressure_rel > 909 and out_temp > -9.8 and winds <= 4.9 and winds > 1 and heat_index <= 0 and out_tmp
Rule 47 for ve (10. 0.7) if pressure_rel > 909 and out_temp > -9.8 and winds <= 4.9 and winds > 1 and heat_index <= 0 and out_tmp
Rule 48 for ve (10. 0.7) if pressure_rel > 909 and out_temp > -9.8 and winds <= 4.9 and winds > 1 and heat_index <= 0 and out_tmp
Rule 49 for ve (10. 0.7) if pressure_rel > 909 and out_temp > -9.8 and winds <= 4.9 and winds > 1 and heat_index <= 0 and out_tmp
Rule 50 for ve (10. 0.7) if pressure_rel > 909 and out_temp > -9.8 and winds <= 4.9 and winds > 1 and heat_index <= 0 and out_tmp
Rule 51 for ve (10. 0.7) if pressure_rel > 909 and out_temp > -9.8 and winds <= 4.9 and winds > 1 and heat_index <= 0 and out_tmp
```

C5.0 rule for high gust

Rule 118 for high gust (8; 0.75)

if Pressure_Rel > 909
and Out_Temp > -9.8
and Winds > 4.9
and Dewp <= 18
and Winds <= 9.9
and Heat_Idx <= 0
and Out_Temp > 2.5
and Winds > 7.3
and Dewp > 8.3
and Out_Temp > 14.9
and Out_Hum <= 90
and Pressure_Rel <= 1018.9
and Winds <= 8.8
and Wind_Dir <= 242
and Out_Hum > 45
and Winds > 8
and Pressure_Rel > 997.7
and Out_Temp <= 24.9
and Out_Hum > 69
and Dewp <= 16.4
and Wind_Dir <= 135
and Out_Hum > 70
and Winds <= 8.3
and Wind_Dir <= 67
and Wind_Dir <= 22
and Out_Hum <= 72
and Pressure_Rel > 1003.1
and Pressure_Rel <= 1010.5
then high gust
C&RT gust prediction
C&RT rules for Gust prediction

  Winds <= 2.9500 [ Ave: 0.346, Effect: -1.122 ] (66,153)
    Winds <= 1.0500 [ Ave: 0.085, Effect: -0.461 ] (50,723)
      Winds <= 0.3500 [ Ave: 0.015, Effect: -0.07 ] (44,789)
        Winds <= 0.0500 [ Ave: 0.002, Effect: -0.013 ] => 0.002 (41,963)
        Winds > 0.0500 [ Ave: 0.209, Effect: 0.194 ] => 0.209 (2,826)
    Winds > 0.3500 [ Ave: 0.611, Effect: 0.526 ] => 0.611 (5,934)
    Winds > 1.0500 [ Ave: 2.061, Effect: 1.515 ] (13,430)
      Winds <= 2.0500 [ Ave: 1.518, Effect: -0.545 ] (6,977)
        Winds <= 1.4500 [ Ave: 1.308, Effect: -0.208 ] => 1.308 (3,717)
        Winds > 1.4500 [ Ave: 1.752, Effect: 0.237 ] => 1.752 (3,260)
      Winds > 2.0500 [ Ave: 2.511, Effect: 0.45 ] (8,453)
        Winds <= 20.9500 [ Ave: 2.379, Effect: -0.132 ] => 2.379 (7,053)
        Winds > 20.9500 [ Ave: 3.177, Effect: 0.686 ] => 3.177 (1,400)
    Winds > 2.9500 [ Ave: 5.453, Effect: 3.688 ] (21,888)
      Winds <= 4.9500 [ Ave: 4.2, Effect: -1.253 ] (13,671)
        Winds <= 3.8500 [ Ave: 3.461, Effect: -0.739 ] (5,968)
          Winds <= 3.5500 [ Ave: 3.18, Effect: -0.28 ] => 3.18 (3,257)
          Winds > 3.5500 [ Ave: 3.797, Effect: 0.337 ] => 3.797 (2,711)
        Winds > 3.8500 [ Ave: 4.773, Effect: 0.573 ] => 4.773 (7,703)
      Winds > 4.9500 [ Ave: 7.538, Effect: 2.085 ] (8,217)
        Wind_Dir <= 0.500 [ Ave: 11.217, Effect: 3.679 ] => 11.217 (1,835)
        Wind_Dir > 0.500 [ Ave: 6.481, Effect: -1.058 ] (6,382)
          Winds <= 15.3500 [ Ave: 5.533, Effect: -0.947 ] => 5.533 (2,473)
          Winds > 15.3500 [ Ave: 7.08, Effect: 0.599 ] => 7.08 (3,909)
        Wind_Dir <= 0.500 [ Ave: 22.322, Effect: 6.228 ] (7,753)
          Winds <= 8.4000 [ Ave: 17.918, Effect: -4.404 ] => 17.918 (2,434)
        Wind_Dir > 0.500 [ Ave: 13.846, Effect: 2.249 ] (21,473)
            Pressure_Rel <= 1004.1500 [ Ave: 14.72, Effect: 3.411 ] => 14.72 (3,325)
          Winds > 10.2500 [ Ave: 16.916, Effect: 3.071 ] (9,715)
C&RT class: error & correct readings
CHAID wind speed <=0

Tuesday 2 December 2013
Wind speed > 13.7
ANN predict gust class

Analysis
- Estimated accuracy: 83.652
- Input Layer: 8 neurons
- Hidden Layer 1: 3 neurons
- Output Layer: 10 neurons
- Relative Importance of Inputs

```
<table>
<thead>
<tr>
<th>Input</th>
<th>Importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Out_Temp</td>
<td>0.340085</td>
</tr>
<tr>
<td>Heat_Indx</td>
<td>0.2521</td>
</tr>
<tr>
<td>Winds</td>
<td>0.224171</td>
</tr>
<tr>
<td>Dwp</td>
<td>0.218885</td>
</tr>
<tr>
<td>Windc</td>
<td>0.186994</td>
</tr>
<tr>
<td>Out_Hum</td>
<td>0.113052</td>
</tr>
<tr>
<td>Pressure_Rel</td>
<td>0.0187872</td>
</tr>
<tr>
<td>Wind_Dir</td>
<td>0.00938279</td>
</tr>
<tr>
<td>Out_Temp</td>
<td>0.340085</td>
</tr>
</tbody>
</table>
```

Fields
- Target: guest_class
- Inputs: Dwp, Heat_Indx, Out_Hum, Out_Temp,
  Pressure_Rel, Wind_Dir, Windc, Winds

Build Settings
- Use partitioned data: false
- Method: Quick
- Stop on: Default
- Set random seed: false
- Prevent overtraining: true
- Sample %: 50.0
- Optimize: Memory

Training Summary
- Model type: Neural net
- Stream: Stream1
- User: sshanmug
- Date built: 23/05/13 01:56
- Application: Clementine 10.1
- Elapsed time for model build: 0 hours, 0 mins, 31 secs

Tuesday 2 December 2013
## Regression

### Variables Entered/Removed(a)

<table>
<thead>
<tr>
<th>Model</th>
<th>Variables Entered</th>
<th>Variables Removed</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Winds, Pressure_Rel, Out_Hum, Wind_Dir, Windc, Out.Temp, Dewp(b)</td>
<td>Enter</td>
<td></td>
</tr>
</tbody>
</table>

a. Dependent Variable: Gust

b. All requested variables entered.

### Model Summary

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.000(a)</td>
<td>.999</td>
<td>.999</td>
<td>5.46882</td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), Winds, Pressure_Rel, Out_Hum, Wind_Dir, Windc, Out.Temp, Dewp

### ANOVA(a)

<table>
<thead>
<tr>
<th>Model</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>4898160057.877</td>
<td>7</td>
<td>699737151.125</td>
<td>23396356.073</td>
<td>.000(b)</td>
</tr>
<tr>
<td>Residual</td>
<td>4239482.448</td>
<td></td>
<td>141751</td>
<td>29.908</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>4902399540.325</td>
<td>14</td>
<td>141758</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. Dependent Variable: Gust

b. Predictors: (Constant), Winds, Pressure_Rel, Out_Hum, Wind_Dir, Windc, Out.Temp, Dewp

### Coefficients(a)

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
<td></td>
</tr>
<tr>
<td>(Constant)</td>
<td>2.671</td>
<td>.127</td>
<td>.000</td>
<td>20.956</td>
</tr>
<tr>
<td>Dewp</td>
<td>8.27E-002</td>
<td>.003</td>
<td>.005</td>
<td>26.786</td>
</tr>
<tr>
<td>Out_Hum</td>
<td>-7.68E-002</td>
<td>.001</td>
<td>-0.007</td>
<td>-75.270</td>
</tr>
<tr>
<td>Out_Temp</td>
<td>1.06E-002</td>
<td>.002</td>
<td>.001</td>
<td>5.748</td>
</tr>
<tr>
<td>Pressure_Rel</td>
<td>4.18E-003</td>
<td>.000</td>
<td>.006</td>
<td>41.164</td>
</tr>
<tr>
<td>Wind_Dir</td>
<td>-3.59E-003</td>
<td>.000</td>
<td>-0.003</td>
<td>-31.307</td>
</tr>
<tr>
<td>Windc</td>
<td>-8.35E-004</td>
<td>.001</td>
<td>.000</td>
<td>-734.463</td>
</tr>
<tr>
<td>Winds</td>
<td>.999</td>
<td>.000</td>
<td>.999</td>
<td>5914.912</td>
</tr>
</tbody>
</table>

a. Dependent Variable: Gust

---

Tuesday 2 December 2013
CHAID

$R$-Gust

Node 0
n 124572
% 100.000
Predicted 6.568

<= 0.000
(0.000, 0.900]
(0.900, 2.400]
(2.400, 4.000]
(4.000, 6.100]
(6.100, 9.000]
(9.000, 13.700]
> 13.700

Node 1
n 41963
% 33.666
Predicted 0.002

Node 18
n 7550
% 6.061
Predicted 0.453

Node 26
n 11930
% 9.577
Predicted 1.663

Node 41
n 13870
% 11.134
Predicted 3.421

Node 51
n 12699
% 10.194
Predicted 6.650

Node 64
n 11896
% 9.549
Predicted 13.022

Node 77
n 12130
% 9.737
Predicted 17.437

Node 91
n 12534
% 10.062
Predicted 23.880
### Factor Analysis

#### Communalities

<table>
<thead>
<tr>
<th></th>
<th>Initial</th>
<th>Extraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pressure_Rel</td>
<td>1.000</td>
<td>.830</td>
</tr>
<tr>
<td>Out.Temp</td>
<td>1.000</td>
<td>.893</td>
</tr>
<tr>
<td>Out.Hum</td>
<td>1.000</td>
<td>.998</td>
</tr>
<tr>
<td>Dewp</td>
<td>1.000</td>
<td>.939</td>
</tr>
<tr>
<td>Windc</td>
<td>1.000</td>
<td>.991</td>
</tr>
<tr>
<td>Winds</td>
<td>1.000</td>
<td>.958</td>
</tr>
<tr>
<td>Wind.Dir</td>
<td>1.000</td>
<td>.999</td>
</tr>
</tbody>
</table>

Extraction Method: Principal Component Analysis.

#### Total Variance Explained

<table>
<thead>
<tr>
<th>Component</th>
<th>Initial Eigenvalues</th>
<th>Extraction Sums of Squared Loadings</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
<td>% of Variance</td>
</tr>
<tr>
<td>1</td>
<td>2.889</td>
<td>41.274</td>
</tr>
<tr>
<td>2</td>
<td>1.295</td>
<td>18.494</td>
</tr>
<tr>
<td>3</td>
<td>1.087</td>
<td>15.527</td>
</tr>
<tr>
<td>4</td>
<td>.920</td>
<td>13.140</td>
</tr>
<tr>
<td>5</td>
<td>.418</td>
<td>5.967</td>
</tr>
<tr>
<td>6</td>
<td>.298</td>
<td>4.254</td>
</tr>
<tr>
<td>7</td>
<td>4.11E-002</td>
<td>1.344</td>
</tr>
</tbody>
</table>

Extraction Method: Principal Component Analysis.

#### Component Matrix (a)

<table>
<thead>
<tr>
<th></th>
<th>Component</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pressure_Rel</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>5</td>
</tr>
<tr>
<td>Out.Temp</td>
<td>- .909</td>
</tr>
<tr>
<td></td>
<td>.75E-003</td>
</tr>
<tr>
<td>Out.Hum</td>
<td>-7.15E-002</td>
</tr>
<tr>
<td></td>
<td>-.744</td>
</tr>
<tr>
<td></td>
<td>-9.50E-002</td>
</tr>
<tr>
<td>Dewp</td>
<td>.909</td>
</tr>
<tr>
<td></td>
<td>-.144</td>
</tr>
<tr>
<td></td>
<td>3.37E-002</td>
</tr>
<tr>
<td>Windc</td>
<td>.663</td>
</tr>
<tr>
<td></td>
<td>.317</td>
</tr>
<tr>
<td>Winds</td>
<td>.300</td>
</tr>
<tr>
<td></td>
<td>7.86E-002</td>
</tr>
<tr>
<td>Wind.Dir</td>
<td>-.204</td>
</tr>
<tr>
<td></td>
<td>-.636</td>
</tr>
<tr>
<td></td>
<td>9.20E-002</td>
</tr>
</tbody>
</table>

Extraction Method: Principal Component Analysis.

a. 5 components extracted.
Conclusions

• Different primary predictors
  – C5.0 => pressure relative
  – C&RT => wind speed
  – CHAID => wind speed
  – Regression test model => wine speed,
    pressure relative, outdoor humidity, wind direction, wind chill, outdoor
    temperature, dew point
  – PCA => pressure relative
    • outdoor temperature, outdoor humidity, dew point, wind chill, w speed,
      w direction

• Future work
  – Deploy online
  – Test other location data
PIXEL CLUSTERING IN SPATIAL DATA MINING
Figure 1  Factors encountered in the grape growing to wine production continuum. Note that viniculture is the study or science of making wines.

viticulture zoning

• Requires extensive knowledge on “Terroir” properties

• Makes it difficult for zoning “new terroirs” / new world wine regions, such as New Zealand, Chilli, South Africa

• New approach is presented → SOM clustering & TDIDT (Top-Down Induction of Decision Trees) using Kumeu wine region, New Zealand

www.cs.utsa.edu/~bylander/cs6243/decision-tree.pdf
background

New Zealand

Kumeu

Kumeu vineyards
methodology

Transform Feature (attribute) layers to raster based
Overlay all raster based attribute layers
Extract pixel based attribute data for NZ vineyards
SOM clustering of vineyard pixel attribute data at different scales
Generate rules for SOM clustering of vineyard pixels using C5.0, CR&T, CHID and QUEST algorithms
Project SOM clustering of vineyard pixel on GIS
SOM clustering

Each neuron $i$ of the SOM $\text{SOM} = a$ weight model vector, $m_i = [m_{i1},\ldots,m_{in}]^T$, neighbourhood $Ni$ of the neuron $i$ distances (such as the similarities) between the vector $x$ and all codebook vectors are computed.

The best matching unit (BMU) $c$, map unit $x$: $||x - m_c|| = \min \{||x - m_i||\}$

The weight vectors are updated

The BMU and its topological neighbours are moved closer to the input vector in the input space.

The update rule for the weight vector of unit $i$ is:

$$m_i(t) + \alpha(t) [x(t) - m_i(t)], iNc(t)$$

$$m_i(t+1) = m_i(t), iNc(t)$$

where $t$ denotes time. $Nc(t)$ is the non-increasing neighbourhood function around the winner unit $c$ and $0 < \alpha(t) < 1$ is a learning coefficient, a decreasing function of time.

Tuesday 2 December 2013
variables used

**Climate variables**

1. Mean annual temperature: strongly influences plant productivity.
2. Mean minimum winter Temperature: influences plant survival.
3. Mean annual solar radiation: determines potential productivity.
4. Monthly water balance ratio: indicates average site “wetness”.
5. Annual water deficit: gives an indication of soil dryness, it is calculated using mean of daily temperature, daily solar radiation and rainfall (Leathwick, Morgan, Wilson, Rutledge, McLeod, & Johnston, 2002)

**Land form variables**

1. Elevation
2. Slope: Major driver of drainage, soil rejuvenation and microclimate
3. Aspect: the direction in which a slope faces
4. Hill shade

**Soil variables**

1. Drainage: influences the oxygen availability in upper soil layers.
2. Acid soluble phosphorous: indicates a key soil nutrient
3. Exchange calcium: both a nutrient and a determinant of soil weathering.
5. Age: separates recent, fertile soils from older less fertile soils.
6. Chemical limitation of plant growth: indicates the presence of salinity of ultramafic substances.
NZ vs Kumeu scale

Figure 1b: SOM cluster profiles, Wat Ba R: monthly water balance ratio
SOM clustering

- Annual solar radiation, annual average and minimum temperatures, acid soluble phosphorous, drainage, elevation, cation exchange, induration, monthly water balance and annual water deficit show similarity.
- Aspect and hill shade show variability that can be used for zoning purposes.
- Age (soil) has one cluster that is 1 year (new fertile) and rest of the clusters are 2 years old (less fertile).
NZ vs Kumeu scale

<table>
<thead>
<tr>
<th>Kumeu C</th>
<th>Freq...</th>
<th>a_sol</th>
<th>a_t...</th>
<th>acid...</th>
<th>age</th>
<th>aspect</th>
<th>che_lim</th>
<th>dra_25</th>
<th>ele_25_26</th>
<th>exch...</th>
<th>hills...</th>
<th>indur...</th>
<th>min...</th>
<th>sl...</th>
<th>w_b...</th>
<th>wa...</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>29.05%</td>
<td>14.90</td>
<td>14.10</td>
<td>1.70</td>
<td>2.00</td>
<td>81.2</td>
<td>1.00</td>
<td>4.50</td>
<td>45.00</td>
<td>1.60</td>
<td>177.7</td>
<td>2.00</td>
<td>4.80</td>
<td>0</td>
<td>2.50</td>
<td>40.16</td>
</tr>
<tr>
<td>2</td>
<td>20.88%</td>
<td>14.90</td>
<td>14.10</td>
<td>1.70</td>
<td>2.00</td>
<td>318.9</td>
<td>1.00</td>
<td>4.60</td>
<td>44.95</td>
<td>1.599</td>
<td>184.4</td>
<td>2.002</td>
<td>4.802</td>
<td>0</td>
<td>2.50</td>
<td>40.16</td>
</tr>
<tr>
<td>3</td>
<td>12.99%</td>
<td>14.90</td>
<td>14.10</td>
<td>1.70</td>
<td>2.00</td>
<td>218.7</td>
<td>1.00</td>
<td>4.60</td>
<td>45.00</td>
<td>1.60</td>
<td>179.5</td>
<td>2.00</td>
<td>4.80</td>
<td>0</td>
<td>2.50</td>
<td>40.16</td>
</tr>
<tr>
<td>4</td>
<td>7.89%</td>
<td>14.80</td>
<td>14.00</td>
<td>1.50</td>
<td>2.00</td>
<td>309.3</td>
<td>1.00</td>
<td>1.50</td>
<td>28.00</td>
<td>2.10</td>
<td>182.3</td>
<td>1.70</td>
<td>4.10</td>
<td>0</td>
<td>2.40</td>
<td>50.42</td>
</tr>
<tr>
<td>5</td>
<td>10.32%</td>
<td>14.90</td>
<td>14.10</td>
<td>1.70</td>
<td>2.00</td>
<td>134.8</td>
<td>1.00</td>
<td>4.50</td>
<td>45.01</td>
<td>1.60</td>
<td>169.4</td>
<td>2.00</td>
<td>4.80</td>
<td>0</td>
<td>2.50</td>
<td>40.16</td>
</tr>
<tr>
<td>6</td>
<td>2.76%</td>
<td>15.00</td>
<td>15.00</td>
<td>1.80</td>
<td>2.00</td>
<td>187.0</td>
<td>1.00</td>
<td>1.90</td>
<td>40.00</td>
<td>2.40</td>
<td>181.2</td>
<td>1.00</td>
<td>6.60</td>
<td>0</td>
<td>2.50</td>
<td>50.28</td>
</tr>
<tr>
<td>7</td>
<td>5.28%</td>
<td>14.80</td>
<td>14.00</td>
<td>1.50</td>
<td>2.00</td>
<td>76.4</td>
<td>1.00</td>
<td>1.50</td>
<td>28.00</td>
<td>2.10</td>
<td>179.3</td>
<td>1.70</td>
<td>4.10</td>
<td>0</td>
<td>2.40</td>
<td>50.42</td>
</tr>
<tr>
<td>8</td>
<td>7.90%</td>
<td>14.90</td>
<td>14.10</td>
<td>1.70</td>
<td>2.00</td>
<td>18.7</td>
<td>1.00</td>
<td>4.60</td>
<td>45.00</td>
<td>1.60</td>
<td>182.6</td>
<td>2.00</td>
<td>4.80</td>
<td>0</td>
<td>2.50</td>
<td>40.16</td>
</tr>
<tr>
<td>9</td>
<td>1.01%</td>
<td>14.90</td>
<td>15.00</td>
<td>2.00</td>
<td>1.001</td>
<td>109.0</td>
<td>1.00</td>
<td>3.899</td>
<td>47.99</td>
<td>2.30</td>
<td>178.9</td>
<td>1.00</td>
<td>6.499</td>
<td>0</td>
<td>2.70</td>
<td>42.96</td>
</tr>
<tr>
<td>10</td>
<td>1.91%</td>
<td>15.00</td>
<td>14.80</td>
<td>1.202</td>
<td>2.00</td>
<td>153.6</td>
<td>1.00</td>
<td>3.103</td>
<td>91.87</td>
<td>1.101</td>
<td>176.6</td>
<td>3.098</td>
<td>6.497</td>
<td>0</td>
<td>2.70</td>
<td>40.44</td>
</tr>
</tbody>
</table>
## CRT & C5 rules

### CRT tree rules created with Kumeu pixels alone show
water deficit > 40.3 and aspect (asp) > 151.99) as
major discerning factor followed by
hill shade/elevation and min temp.
A temp has been used in 2 rules

<table>
<thead>
<tr>
<th>Rule no</th>
<th>Instance; confidence</th>
<th>Rule asp; aspect, hs; hill shade, wd; water deficit, ele; elevation 25m resolution</th>
<th>SOM cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 46; 1.0</td>
<td>if wd&lt;= 40.16 and asp &lt;= 106.56 and &gt; 29.15 and hs &lt;= 173 and &gt; 172</td>
<td>one</td>
<td></td>
</tr>
<tr>
<td>2 309; 1.0</td>
<td>if wd&lt;= 40.16 and asp &lt;= 136.38 and &gt; 29.15 and hs &lt;= 175 and &gt; 173</td>
<td>one</td>
<td></td>
</tr>
<tr>
<td>3 5; 1.0</td>
<td>if wd&lt;= 40.16 and asp &lt;= 145.11 and &gt; 136.38 and hs &lt;= 174 and &gt; 173</td>
<td>one</td>
<td></td>
</tr>
<tr>
<td>4 1,916; 1.0</td>
<td>if wd&lt;= 40.16 and asp &lt;= 151.34 and &gt; 29.15 and hs &lt;= 180 and &gt; 175</td>
<td>one</td>
<td></td>
</tr>
<tr>
<td>5 4; 1.0</td>
<td>if wd&lt;= 40.16 and asp &lt;= 156.37 and &gt; 151.34 and hs &lt;= 176 and &gt; 175</td>
<td>one</td>
<td></td>
</tr>
<tr>
<td>6 2; 1.0</td>
<td>if wd&lt;= 40.16 and asp &lt;= 154.45 and &gt; 151.34 and ele_25 in [45] and hs &lt;= 177 and &gt; 176</td>
<td>one</td>
<td></td>
</tr>
<tr>
<td>1 2; 1.0</td>
<td>if wd&lt;= 40.16 and asp &lt;= 277.27 and &gt; 151.34 and ele_25 in [0] and hs &lt;= 182 and &gt; 176</td>
<td>two</td>
<td></td>
</tr>
<tr>
<td>2 7; 1.0</td>
<td>if wd&lt;= 40.16 and asp &lt;= 277.27 and &gt; 264.29 and ele_25 in [45] and hs &lt;= 182 and &gt; 181</td>
<td>two</td>
<td></td>
</tr>
<tr>
<td>3 126; 1.0</td>
<td>if wd&lt;= 40.16 and asp &lt;= 277.27 and &gt; 176 and hs &gt; 182</td>
<td>two</td>
<td></td>
</tr>
<tr>
<td>4 63; 0.984</td>
<td>if wd&lt;= 40.16 and asp &lt;= 284.39 and &gt; 277.27 and hs &gt; 180</td>
<td>two</td>
<td></td>
</tr>
<tr>
<td>5 1,425; 1.0</td>
<td>if wd&lt;= 40.16 and asp &gt; 284.39</td>
<td>two</td>
<td></td>
</tr>
<tr>
<td>1 13; 0.923</td>
<td>if wd&lt;= 40.16 and asp &lt;= 154.45 and &gt; 151.34 and ele_25 in [45] and hs &lt;= 181 and &gt; 177</td>
<td>three</td>
<td></td>
</tr>
<tr>
<td>2 958; 1.0</td>
<td>if wd&lt;= 40.16 and asp &lt;= 277.27 and &gt; 154.45 and ele_25 in [45] and hs &lt;= 181 and &gt; 176</td>
<td>three</td>
<td></td>
</tr>
<tr>
<td>3 39; 1.0</td>
<td>if wd&lt;= 40.16 and asp &lt;= 264.29 and &gt; 151.34 and ele_25 in [45] and hs &lt;= 182 and &gt; 181</td>
<td>three</td>
<td></td>
</tr>
<tr>
<td>4 8; 1.0</td>
<td>if wd&lt;= 40.16 and asp &lt;= 284.39 and &gt; 277.27 and hs &lt;= 180</td>
<td>three</td>
<td></td>
</tr>
<tr>
<td>1 620; 1.0</td>
<td>if wd&gt; 40.16 and asp &gt; 190.35 and min_temp &lt;= 4.8</td>
<td>four</td>
<td></td>
</tr>
<tr>
<td>1 484; 1.0</td>
<td>if wd&gt; 40.16 and asp &lt;= 151.34 and hs &lt;= 172</td>
<td>five</td>
<td></td>
</tr>
<tr>
<td>2 22; 1.0</td>
<td>if wd&gt; 40.16 and asp &lt;= 136.38 and &gt; 106.56 and hs &lt;= 173 and &gt; 172</td>
<td>five</td>
<td></td>
</tr>
<tr>
<td>3 11; 1.0</td>
<td>if wd&gt; 40.16 and asp &lt;= 151.34 and &gt; 136.38 and hs &lt;= 173 and &gt; 172</td>
<td>five</td>
<td></td>
</tr>
<tr>
<td>4 9; 1.0</td>
<td>if wd&gt; 40.16 and asp &lt;= 151.34 and &gt; 145.11 and hs &lt;= 174 and &gt; 173</td>
<td>five</td>
<td></td>
</tr>
<tr>
<td>5 10; 1.0</td>
<td>if wd&gt; 40.16 and asp &lt;= 151.34 and &gt; 136.38 and hs &lt;= 175 and &gt; 174</td>
<td>five</td>
<td></td>
</tr>
<tr>
<td>6 22; 1.0</td>
<td>if wd&gt; 40.16 and asp &lt;= 156.37 and &gt; 151.34 and hs &lt;= 175</td>
<td>five</td>
<td></td>
</tr>
<tr>
<td>7 253; 1.0</td>
<td>if wd&gt; 40.16 and asp &lt;= 277.27 and &gt; 156.37 and hs &lt;= 176</td>
<td>five</td>
<td></td>
</tr>
<tr>
<td>1 217; 1.0</td>
<td>if wd&gt; 40.16 and min_temp &gt; 6.5</td>
<td>six</td>
<td></td>
</tr>
<tr>
<td>1 415; 1.0</td>
<td>if wd&gt; 40.16 and asp &lt;= 190.35 and min_temp &lt;= 4.8</td>
<td>seven</td>
<td></td>
</tr>
<tr>
<td>1 52; 1.0</td>
<td>if wd&gt; 40.16 and asp &lt;= 29.15 and hs &lt;= 180 and hs &gt; 172</td>
<td>eight</td>
<td></td>
</tr>
<tr>
<td>2 591; 1.0</td>
<td>if wd&gt; 40.16 and asp &lt;= 151.34 and hs &gt; 180</td>
<td>eight</td>
<td></td>
</tr>
<tr>
<td>1 79; 1.0</td>
<td>if wd&gt; 40.16 and a_temp &gt; 14.8 and min_temp &lt;= 6.5 and &gt; 4.8</td>
<td>nine</td>
<td></td>
</tr>
<tr>
<td>1 150; 1.0</td>
<td>if wd&gt; 40.16 and a_temp &lt;= 14.8 and min_temp &gt;= 6.5 and &gt; 4.8</td>
<td>ten</td>
<td></td>
</tr>
</tbody>
</table>
water deficit in Kumeu

water deficit > 40.16 or <= 40.16
CHAID tree and rules split elevation into 5 classes (=0/=40, =28, =40, =48 and =92)
CHAID tree
Ele = 0.45
CHAIID rules

\text{sod} = 0 \text{ or } \text{sod} = 45 \text{ [Mode: one] } (6,377)
\text{asp} \leq 31.7900 \text{ [Mode: eight] } (585)
\text{hs} \leq 180 \text{ [Mode: eight] } \Rightarrow \text{eight} \ (63; 0.825)
\text{hs} > 180 \text{ [Mode: eight] } \Rightarrow \text{eight} \ (522; 1.0)
\text{asp} > 31.7900 \text{ and } \text{asp} \leq 57.9900 \text{ [Mode: one] } (684)
\text{hs} \leq 180 \text{ [Mode: one] } \Rightarrow \text{one} \ (616; 1.0)
\text{hs} > 180 \text{ [Mode: eight] } \Rightarrow \text{eight} \ (68; 1.0)
\text{asp} > 57.9900 \text{ and } \text{asp} \leq 84.9900 \text{ [Mode: one] } (715)
\text{hs} \leq 180 \text{ [Mode: one] } \Rightarrow \text{one} \ (714)
\text{hs} \leq 173 \text{ [Mode: five] } \Rightarrow \text{five} \ (50; 0.7)
\text{hs} > 173 \text{ [Mode: one] } \Rightarrow \text{one} \ (664; 1.0)
\text{hs} > 180 \text{ [Mode: eight] } \Rightarrow \text{eight} \ (1, 1.0)
\text{asp} > 84.9900 \text{ and } \text{asp} \leq 107.4800 \text{ [Mode: one] } (680)
\text{hs} \leq 173 \text{ [Mode: five] } \Rightarrow \text{five} \ (209; 0.852)
\text{hs} > 173 \text{ [Mode: one] } \Rightarrow \text{one} \ (471; 1.0)
\text{asp} > 107.4800 \text{ and } \text{asp} \leq 139.5900 \text{ [Mode: one] } (675)
\text{hs} \leq 173 \text{ [Mode: five] } \Rightarrow \text{five} \ (255; 1.0)
\text{hs} > 173 \text{ and } \text{hs} \leq 176 \text{ [Mode: one] } \Rightarrow \text{one} \ (177; 0.977)
\text{hs} > 176 \text{ [Mode: one] } \Rightarrow \text{one} \ (243; 1.0)
\text{asp} > 139.5900 \text{ and } \text{asp} \leq 203.9300 \text{ [Mode: five] } (690)
\text{hs} \leq 173 \text{ [Mode: five] } \Rightarrow \text{five} \ (198; 1.0)
\text{hs} > 173 \text{ and } \text{hs} \leq 176 \text{ [Mode: five] } \Rightarrow \text{five} \ (150; 0.853)
\text{hs} > 176 \text{ and } \text{hs} \leq 178 \text{ [Mode: three] } (237)
\text{sp} \leq 0.06 \text{ [Mode: three] } \Rightarrow \text{three} \ (156; 0.987)
\text{sp} > 0.06 \text{ [Mode: three] } \Rightarrow \text{three} \ (81; 0.815)
\text{hs} > 178 \text{ [Mode: three] } \Rightarrow \text{three} \ (105; 0.81)
\text{asp} > 203.9300 \text{ and } \text{asp} \leq 262.1000 \text{ [Mode: three] } (686)
\text{hs} \leq 182 \text{ [Mode: three] } (633)
\text{a_temp} \leq 14.1 \text{ [Mode: three] } (632)
\text{hs} > 176 \text{ [Mode: five] } \Rightarrow \text{five} \ (53; 1.0)
\text{hs} > 176 \text{ [Mode: three] } \Rightarrow \text{three} \ (619; 1.0)
\text{a_temp} > 14.1 \text{ [Mode: two] } \Rightarrow \text{two} \ (1; 1.0)
\text{hs} > 182 \text{ [Mode: two] } \Rightarrow \text{two} \ (53; 1.0)
\text{asp} > 262.1000 \text{ and } \text{asp} \leq 307.7300 \text{ [Mode: two] } (553)
\text{hs} \leq 181 \text{ [Mode: three] } \Rightarrow \text{three} \ (154; 0.604)
\text{hs} > 181 \text{ [Mode: two] } \Rightarrow \text{two} \ (399; 0.997)
\text{asp} > 307.7300 \text{ [Mode: two] } \Rightarrow \text{two} \ (1, 109; 1.0)

\text{sod} = 28 \text{ [Mode: four] } (1,035)
\text{asp} \leq 203.9300 \text{ [Mode: seven] } (418)
\text{asp} \leq 139.5900 \text{ [Mode: seven] } \Rightarrow \text{seven} \ (353; 1.0)
\text{asp} > 139.5900 \text{ [Mode: seven] } \Rightarrow \text{seven} \ (65; 0.954)
\text{asp} > 203.9300 \text{ [Mode: four] } \Rightarrow \text{four} \ (617; 1.0)

\text{sod} = 40 \text{ [Mode: six] } \Rightarrow \text{six} \ (217; 1.0)
\text{sod} = 48 \text{ [Mode: nine] } \Rightarrow \text{nine} \ (79; 1.0)
\text{sod} = 92 \text{ [Mode: ten] } \Rightarrow \text{ten} \ (150; 1.0)

Aspect and hillshade also used in the rules
for clusters three, five and two annual average temperature is used (in italics).
SOM clusters six, nine and ten are defined purely on elevation with 217, 79 with 150 instances respectively all at 100% confidence.
Clusters seven and four vary in elevation and aspect.
Quest tree rules

- ele_25 = 0 or ele_25 = 45 or ele_25 = 92 [Mode: one]
  - ele_25 = 92 [Mode: ten] -> ten
- ele_25 = 0 or ele_25 = 45 [Mode: one]
  - aspect <= 177.7973 [Mode: one]
    - hillshad <= 172.589 [Mode: five] -> five
    - hillshad > 172.589 [Mode: one]
      - aspect <= 143.0628 [Mode: one] -> one
      - aspect > 143.0628 [Mode: three] -> three
  - aspect > 177.7973 [Mode: two]
    - aspect <= 267.2804 [Mode: three] -> three
    - aspect > 267.2804 [Mode: two] -> two
- ele_25 = 28 or ele_25 = 40 or ele_25 = 48 [Mode: four]
  - ele_25 = 28 [Mode: four]
    - aspect <= 216.3665 [Mode: seven] -> seven
    - aspect > 216.3665 [Mode: four] -> four
  - ele_25 = 40 or ele_25 = 48 [Mode: six]
    - min_temp <= 6.5000 [Mode: nine] -> nine
    - min_temp > 6.5000 [Mode: six] -> six
# Regression

## Variables Entered/Removed

<table>
<thead>
<tr>
<th>Model</th>
<th>Variables Entered</th>
<th>Variables Removed</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>wa_defi, age, hillshad, slope, aspect, min_temp, acid_s_p, induration(b)</td>
<td></td>
<td>Enter</td>
</tr>
</tbody>
</table>

a. Dependent Variable: Rno  
b. Tolerance = .000 limits reached.

## Model Summary

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.638(a)</td>
<td>.407</td>
<td>.407</td>
<td>1.90878</td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), wa_defi, age, hillshad, slope, aspect, min_temp, acid_s_p, induration

## ANOVA

<table>
<thead>
<tr>
<th>Model</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>19656.497</td>
<td>8</td>
<td>2457.062</td>
<td>674.379</td>
<td>000(b)</td>
</tr>
<tr>
<td>Residual</td>
<td>28597.388</td>
<td>7849</td>
<td>3.643</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>48253.884</td>
<td>7857</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. Dependent Variable: Rno  
b. Predictors: (Constant), wa_defi, age, hillshad, slope, aspect, min_temp, acid_s_p, induration

## Coefficients

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
<td></td>
</tr>
<tr>
<td>(Constant)</td>
<td>-6.480</td>
<td>2.014</td>
<td>-3.218</td>
<td>001</td>
</tr>
<tr>
<td>aspect</td>
<td>-5.89E-03</td>
<td>.000</td>
<td>-.270</td>
<td>28.249</td>
</tr>
<tr>
<td>min_temp</td>
<td>1.504</td>
<td>.050</td>
<td>.297</td>
<td>30.117</td>
</tr>
<tr>
<td>hillshad</td>
<td>8.71E-03</td>
<td>.005</td>
<td>.018</td>
<td>1.952</td>
</tr>
<tr>
<td>induration</td>
<td>2.539</td>
<td>.217</td>
<td>.272</td>
<td>11.713</td>
</tr>
<tr>
<td>acid_s_p</td>
<td>-3.624</td>
<td>.463</td>
<td>-.157</td>
<td>7.826</td>
</tr>
<tr>
<td>age</td>
<td>-5.875</td>
<td>.255</td>
<td>-.237</td>
<td>23.083</td>
</tr>
<tr>
<td>slope</td>
<td>-2.852</td>
<td>.425</td>
<td>-.061</td>
<td>6.703</td>
</tr>
<tr>
<td>wa_defi</td>
<td>.360</td>
<td>.016</td>
<td>.544</td>
<td>22.101</td>
</tr>
</tbody>
</table>

a. Dependent Variable: Rno
ANN test results

Analysis

Estimated accuracy: 92.861
Input Layer: 20 neurons
Hidden Layer 1: 3 neurons
Output Layer: 10 neurons

Relative Importance of Inputs

<table>
<thead>
<tr>
<th>Field</th>
<th>Importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>aspect</td>
<td>0.335183</td>
</tr>
<tr>
<td>hillshad</td>
<td>0.30243</td>
</tr>
<tr>
<td>ele_25</td>
<td>0.242831</td>
</tr>
<tr>
<td>age</td>
<td>0.145853</td>
</tr>
<tr>
<td>dra_25</td>
<td>0.101321</td>
</tr>
<tr>
<td>wa_defi</td>
<td>0.0725418</td>
</tr>
<tr>
<td>induration</td>
<td>0.0692616</td>
</tr>
<tr>
<td>a_temp</td>
<td>0.0647437</td>
</tr>
<tr>
<td>min_temp</td>
<td>0.0521352</td>
</tr>
<tr>
<td>a_sol</td>
<td>0.0475153</td>
</tr>
<tr>
<td>w_bal_ra</td>
<td>0.0293795</td>
</tr>
<tr>
<td>exch_cal</td>
<td>0.0234538</td>
</tr>
<tr>
<td>acid_s_p</td>
<td>0.0108849</td>
</tr>
<tr>
<td>slope</td>
<td>0.00307043</td>
</tr>
</tbody>
</table>

Fields

Target: Rno

Build Settings

Use partitioned data: false
Method: Quick
Stop on: Default
Set random seed: false
Prevent overtraining: true
Sample %: 50.0
Optimize: Memory

Training Summary

Model type: Neural net
Stream: Stream1
User: sshanmug
Date built: 25/06/13 23:24
Application: Clementine 10.1
Elapsed time for model build: 0 hours, 0 mins, 3 secs

Tuesday 2 December 2013
conclusions

• Traditional approaches to zoning require extensive knowledge → makes zoning new areas/ “terroirs” impossible
• It's possible to identify main/major contributory attributes using even low res thematic maps
• SOM and TDIDT approach gives a means overcome this
• For Kumeu
  – Water deficit, elevation (along with aspect and hill shade)
  – annual minimum and average temperatures (same as New Zealand regional/ macro-scale GDD)
• Regression test results: water deficit, age, hill shade, slope, aspect, min temp, acid sol phosphorous, induration as predictors with .407 adjusted $R^2$