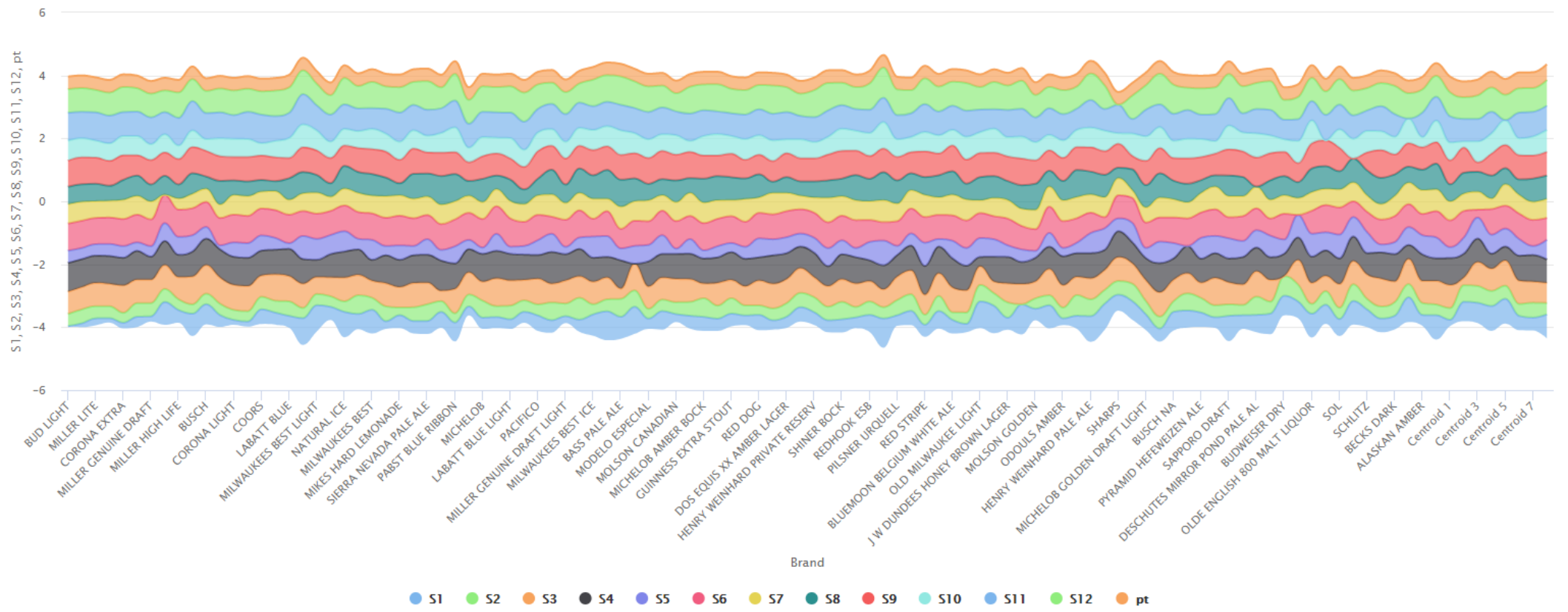


Segmentation and Analysis of Brand Behavioural Patterns – The Preliminary Results

Application of Unsupervised Shapelet Discovery to Sales Data

Jacob L. Cybulski, DISBA, Deakin University
Van Nguyen, DISBA, Deakin University
Chris Dubelaar, Marketing, Deakin University

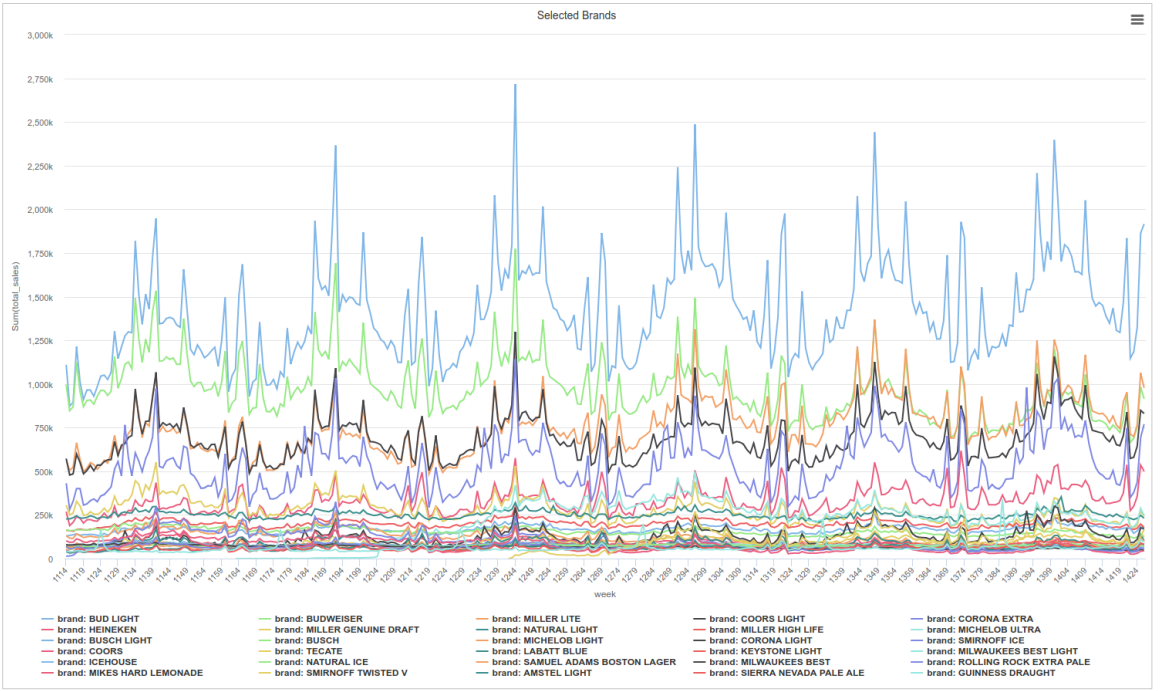
Seminar Plan

- Introduction
- Problem and data
- What are shapelets
- Applications of shapelets
- Shapelet extraction and matching
- Shapelet clustering and diagnostics
- Brands behavioural analysis
- Competitors
- Summary and references
- Related projects

DEAKIN UNIVERSITY

Introduction

- Organisations often struggle understanding their “place” in the marketplace
- Beyond direct competitors, it is not easy to identify “like” companies
- Organisational similarities are easy to detect when analysing their explicit characteristics

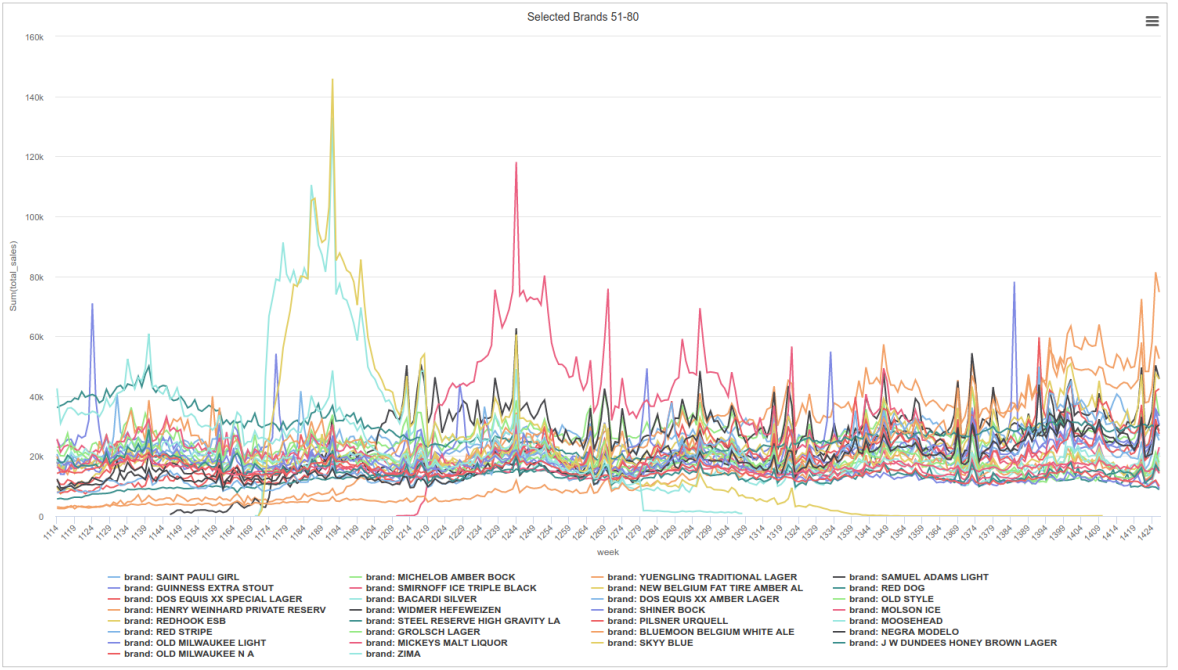


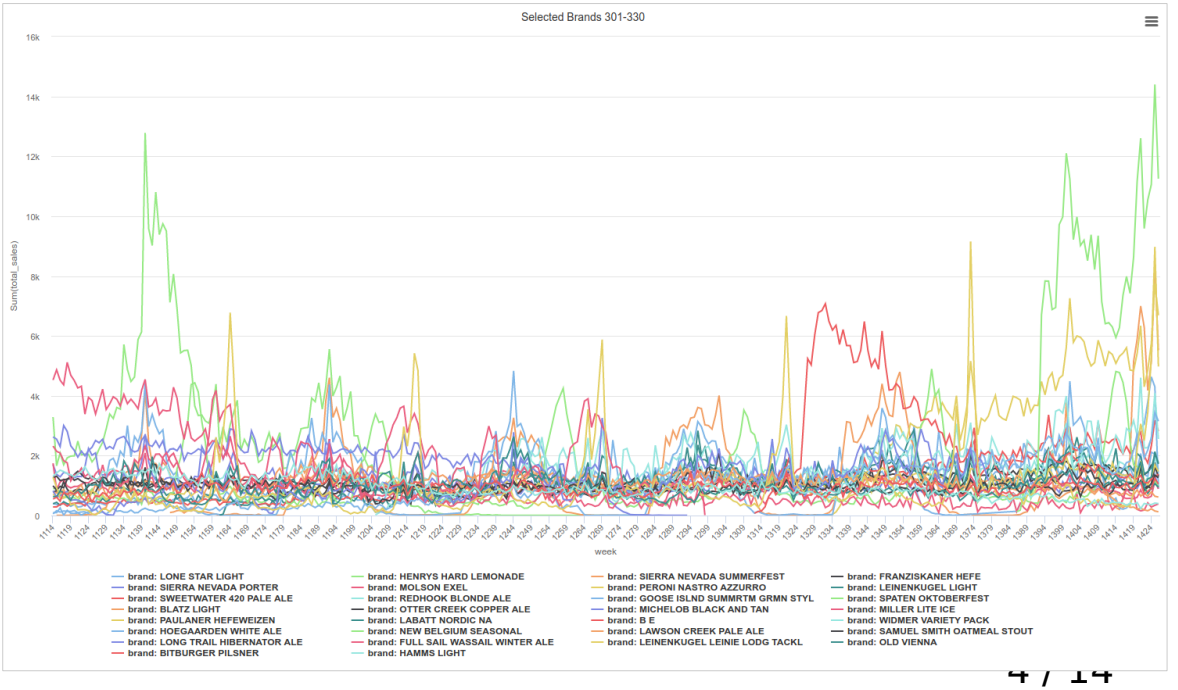
- Similarity of their behaviour, however, is difficult to analyse and interpret
- This project therefore aims to analyse behavioural patterns of a business from its sales data
- Specifically to identify similarities of beer brands (USA) based on their weekly sales

DEAKIN UNIVERSITY

IRI Data

- American market research company IRI collects sales data across the USA
- The project used the IRI data set covering 6 years of sales from 2001 to 2006 across stores, markets, and including panel data
- The IRI subset contained weekly sales of top 100 brands over the time period of 313 weeks
- Top brands seemed to act in unison, the lower-ranking brands showed increasingly erratic sales

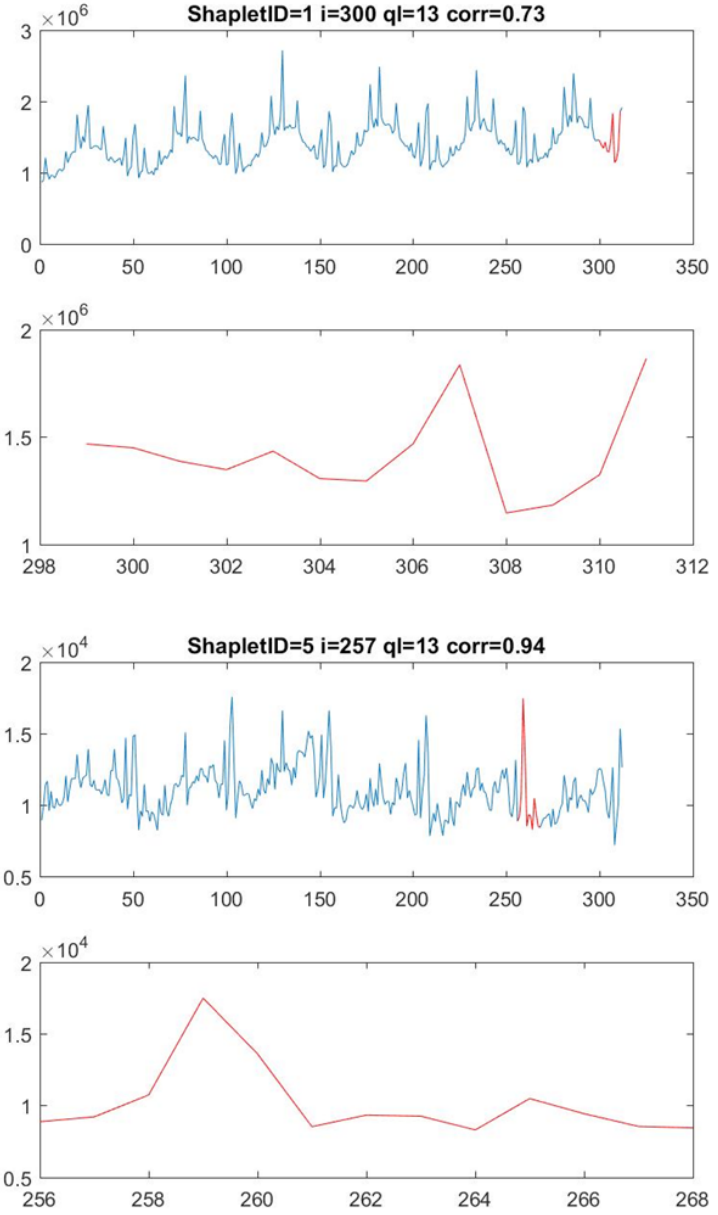




DEAKIN UNIVERSITY

Method: *Shapelets*

- The analysis of IRI data relied on *dynamic shapelet discovery* a method of time series analysis
- *Shapelets* are distinguishing segments of time series data (Grabocka et al 2015, 2016)
- Similar concepts: *dynamic time warping* (time stretch) and *motifs* (discrete)
- The novelty of this work is primarily its business context
- Shapelets are most commonly used for data classification
- *U-shapelets* are unsupervised shapelets used mainly for time series clustering (as used here)

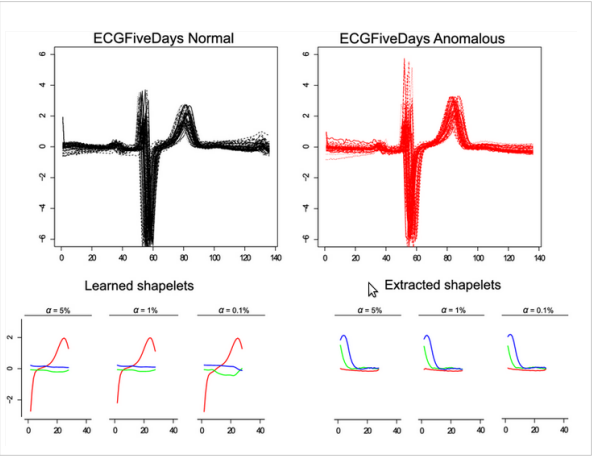


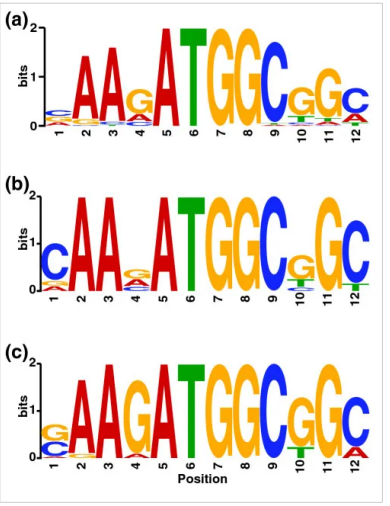
Sample time series and their selected shapelets

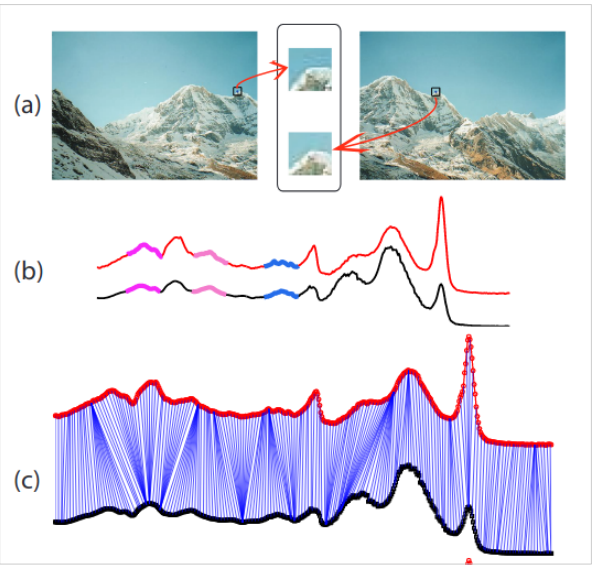
DEAKIN UNIVERSITY

Applications

- Typical applications of shapelets / motifs are in signal processing, e.g. *ECG analysis* to detect heart attacks (Beggel et al 2019)
- Other applications include *genetic sequence matching* (Whitfield et al 2012) or *image analysis* (Zhao & Itti 2016)
- There are an increasing number of business applications, e.g. in the *analysis of stock trading* (Kim et al 2018)
- This project aims at *analysis of product sales* grouped by types, brands and vendors







DEAKIN UNIVERSITY

Method

The shapelets method includes the following (Zakaria et al 2012):

- Calculation of the distance between two time series – here Euclidean distance
- Gap calculation between the sub-sequences of a series resulting from the introduction of a new shapelet
- An algorithm extracting u-shapelets that have the maximum gap, i.e. discriminating abilities
- Clustering of a time series using shapelets and optimised using a Rand Index (CRI)

Algorithm 1. *U-shapelets Extraction*

Input:
- *Dataset D of N (=100) time-series*
- *Desired length of u-shapelets is ql (=13 weeks=1 quarter)*

Output:
- *Set of extracted u-shapelets*

for *each time-series subsequence of length t:*
  compute the Gap *between distances between the subsequence and each of two subsets*
    if *the Gap is maximum:*
      Check the ***discriminating ability*** *of that subsequence*
      Subsequence is ***added as u-shapelet***
    Remove similar time-series *that have distance less than threshold Θ*
return *set of extracted u-shapelets*

Algorithm 2. *Clustering using Distance Map*

Input:
- *Time-series T*
- *Set of extracted u-shapelets S'*
- *Number of clusters k (random number, then adjust after optimization)*

Output:
- *Class labels, Distance Map, CRI values*

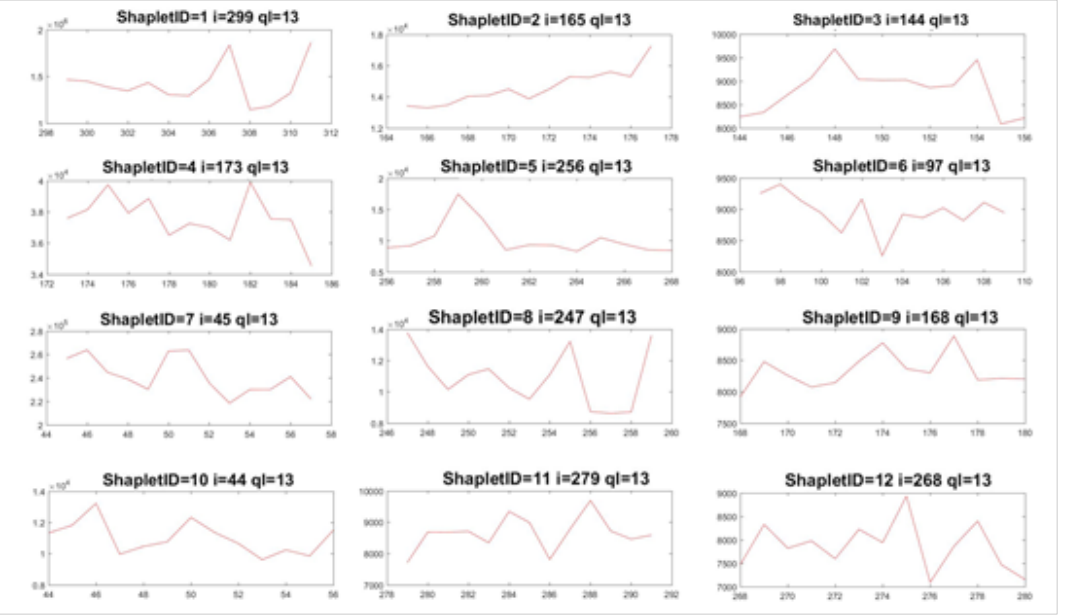
for *each extracted u-shapelet S'*
  compute distances *towards all time-series T, i.e. sdist(S', T)*
  add these distances to the ***Distance Map***
repeat *n times to avoid poor clustering*
  run k-means clustering (based on existing ***Distance Map****)*
  if *clustering results in* ***minimum point-to-centroid distance***
    get the ***Cluster Labels***
    compute ***CRI*** *of clustering using existing u-shapelets*
return *Cluster Labels, Distance Map, CRI values*

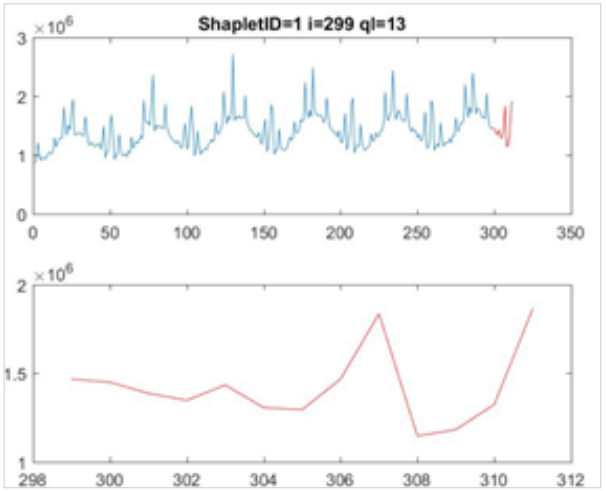
DEAKIN UNIVERSITY

Process

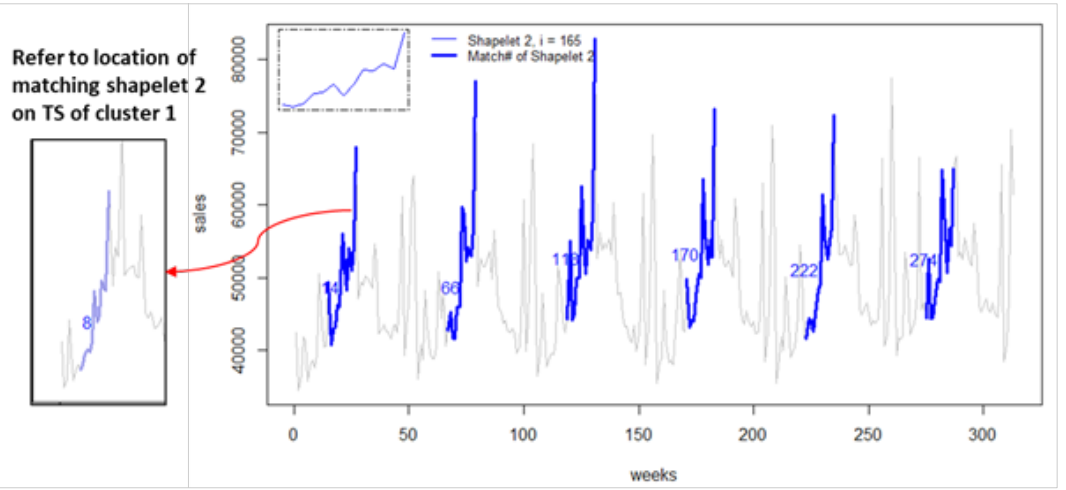
- Selected 100 brands and their time series sales data
- Extracted 12 shapelets from the selected time series
- The shapelets were then be matched against each time series to identify their best matching sub-sequence
- The matching process was then repeated to address seasonality

Shapelets extracted from the time series data set





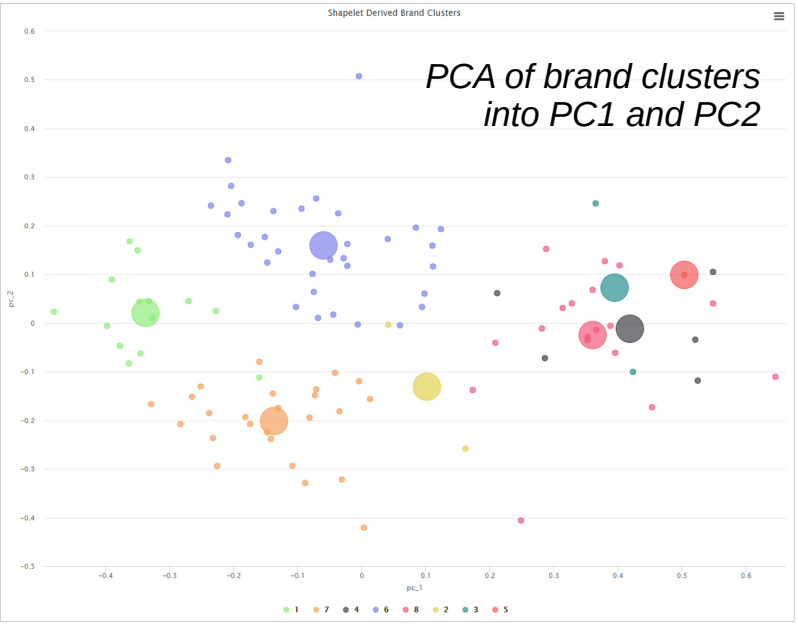
Shapelet matched against time series



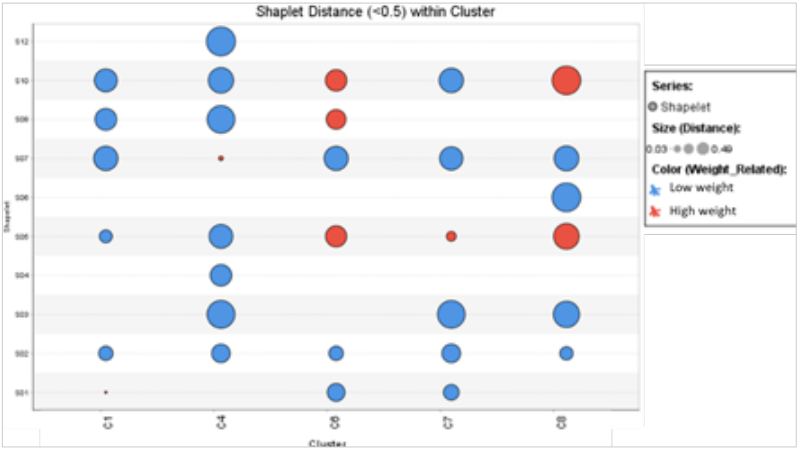
Shapelet repeatedly matched against a time series

Clustering

- Distances between each time series (brand sales) have been measured against each shapelet
- This resulted in a matrix (see below) where each brand was an example set described by variables derived from shapelets
- The new example set could then be used for data clustering and anomaly detection
- To determine the importance of each u-shapelet in distinguishing one cluster from the others, we measured the weight of each u-shapelet with respect to the cluster by using information gain. The higher the information gain of a shapelet, the more relevant it is in forming the cluster.



Shapelet significance for cluster formation



Brand x shapelets matrix

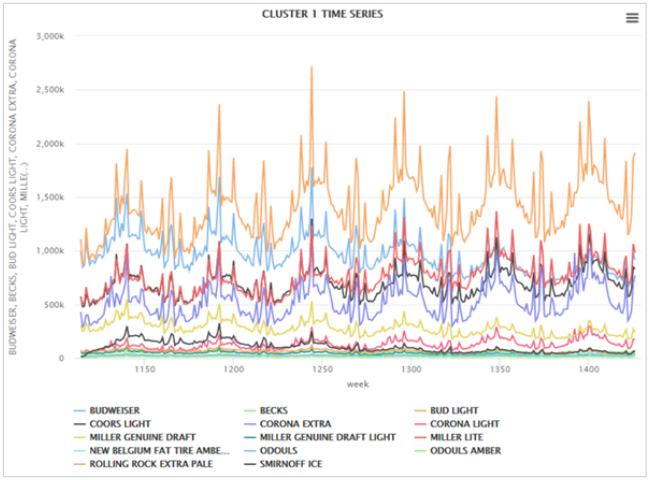
Brand ↑	Cluster	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12
ALASKAN AMBER	6	0.346	0.380	0.700	0.845	0.611	0.678	0.752	0.652	0.714	0.325	0.779	0.734
AMSTEL LIGHT	7	0.392	0.438	0.771	0.885	0.293	0.881	0.616	0.795	0.805	0.579	0.808	0.731
BASS PALE ALE	8	0.700	0.570	0.326	0.900	0.378	0.624	0.701	0.858	0.793	0.776	0.811	0.903
BECKS	1	0.287	0.377	0.705	0.739	0.290	0.863	0.542	0.468	0.596	0.464	0.760	0.758
BECKS DARK	7	0.434	0.433	0.656	0.870	0.381	0.829	0.636	0.687	0.631	0.563	0.721	0.897
BLUEMOON BELGIUM WHITE ALE	6	0.436	0.240	0.787	0.927	0.560	0.809	0.656	0.748	0.812	0.501	0.769	0.761
BUD ICE	6	0.591	0.366	0.699	0.751	0.597	0.779	0.492	0.507	0.903	0.529	0.851	0.692
BUD ICE LIGHT	2	0.609	0.431	0.482	0.814	0.574	0.867	0.515	0.719	0.823	0	0.808	0.745

DEAKIN UNIVERSITY

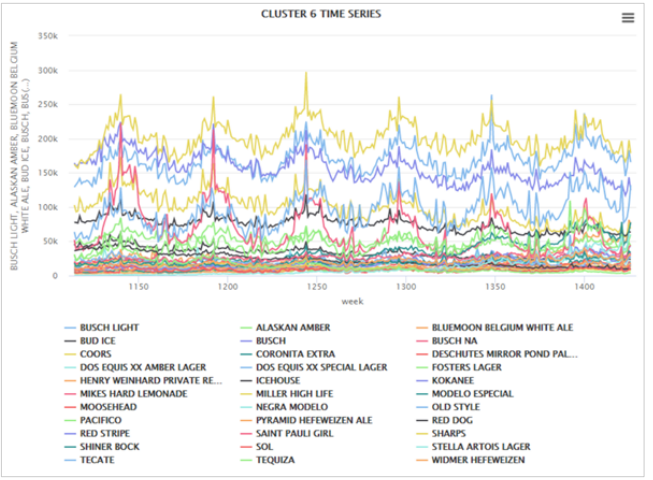
Cluster Diagnostics

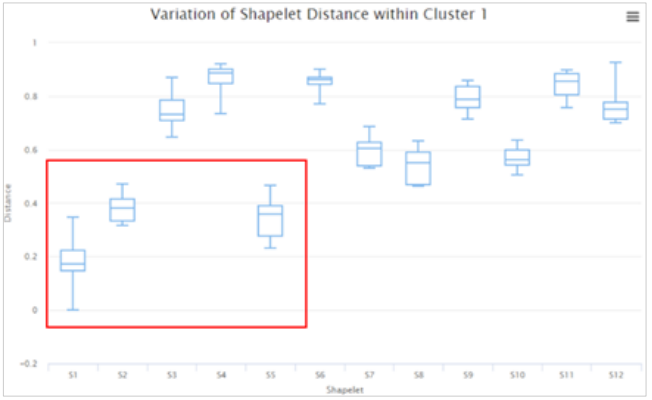
- Clusters should be consisting of similar members, i.e. be cohesive
- Clusters should also be distinct one from the others
- Some clusters were highly cohesive as they included brands of near-identical behaviour (cluster 1)
- Others consisted of brands having a greater variety of behavioural patterns (cluster 6)

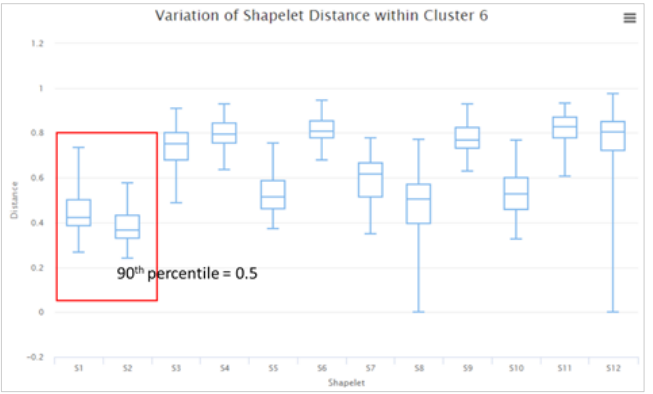
Cluster 1 and its members



Cluster 6 and its members





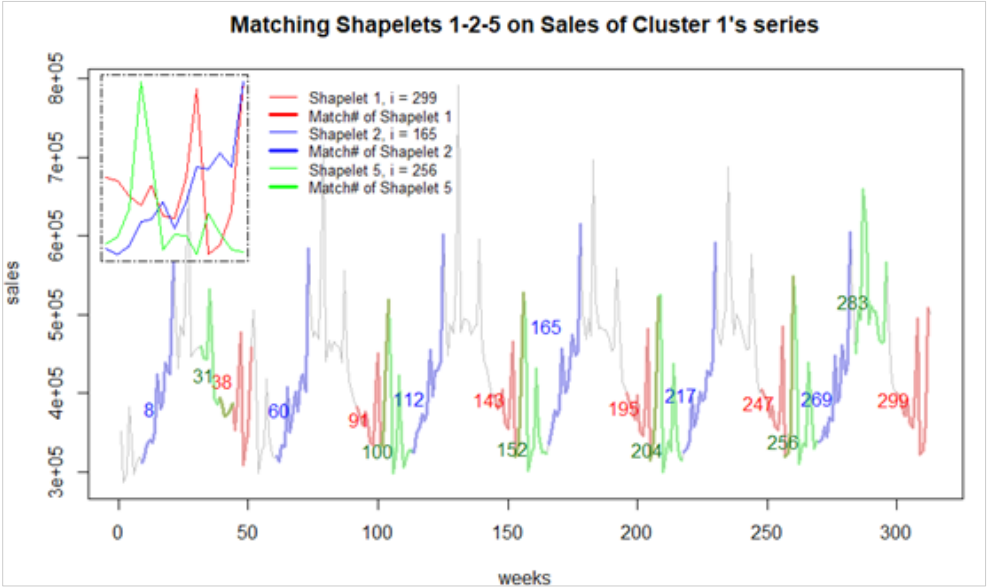


Behavioural similarity were noted in spite of very different volume of sales, which indicates that all beer vendors representing their brands respond to the identical market conditions and events and are thus *direct competitors* !

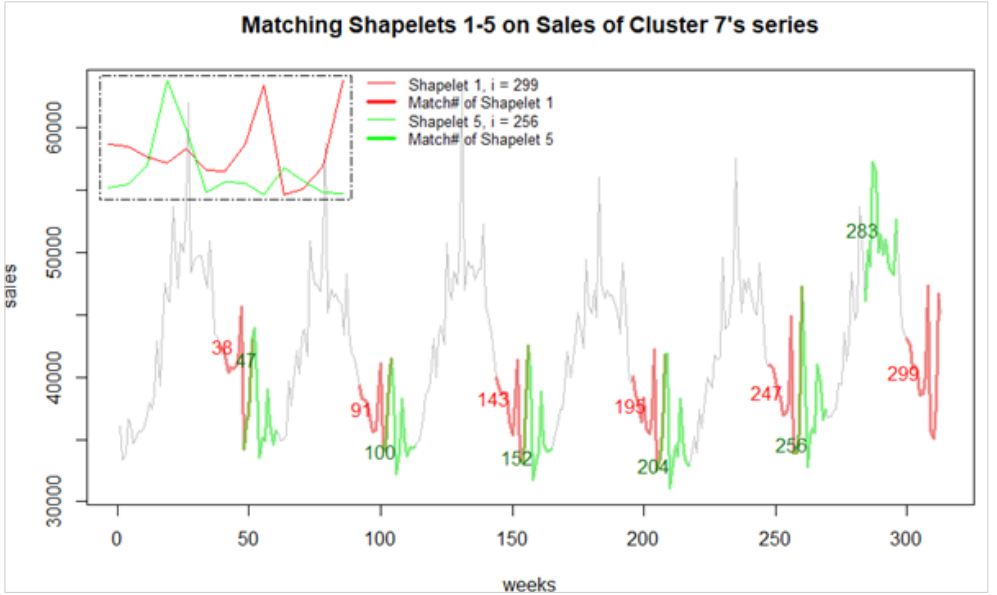
Behavioural Analysis

- Brand behavioural patterns can be analysed in terms of shapelets present in their time series, either individually or by cluster
- The cluster centroid (average of cluster members) represents an “ideal” cluster member
- Here we can see two cluster centroids 1 and 7 with their most significant shapelets highlighted
- Both clusters share some similarities described by common shapelets 1 and 5 present in their time series
- However cluster 1 has an extra very distinct feature in terms of shapelet 2, which cannot be well identified in cluster 7

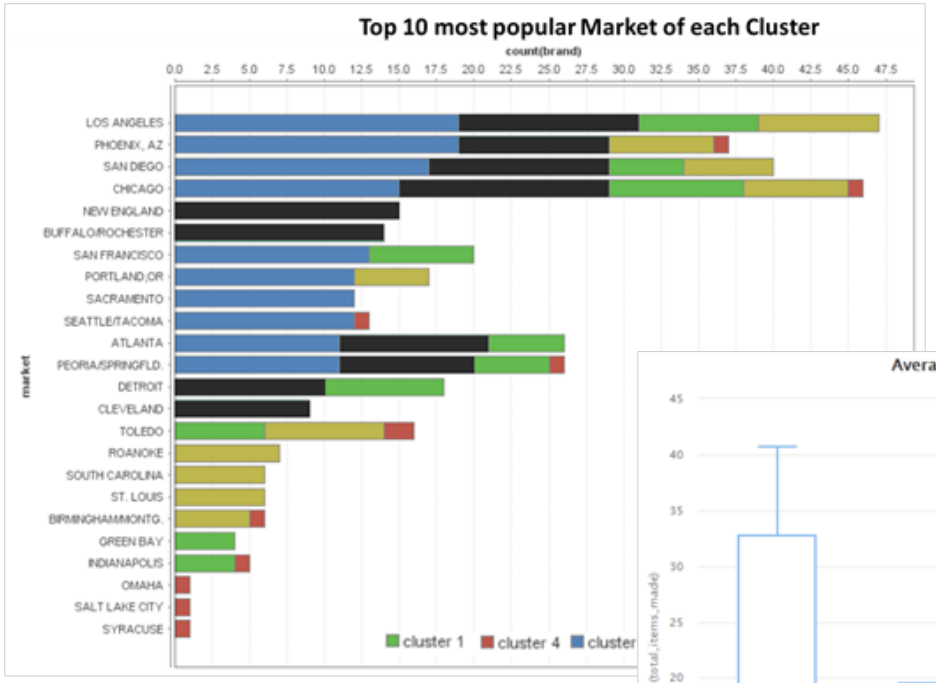
Centroid of cluster 1 and its behavioural patterns

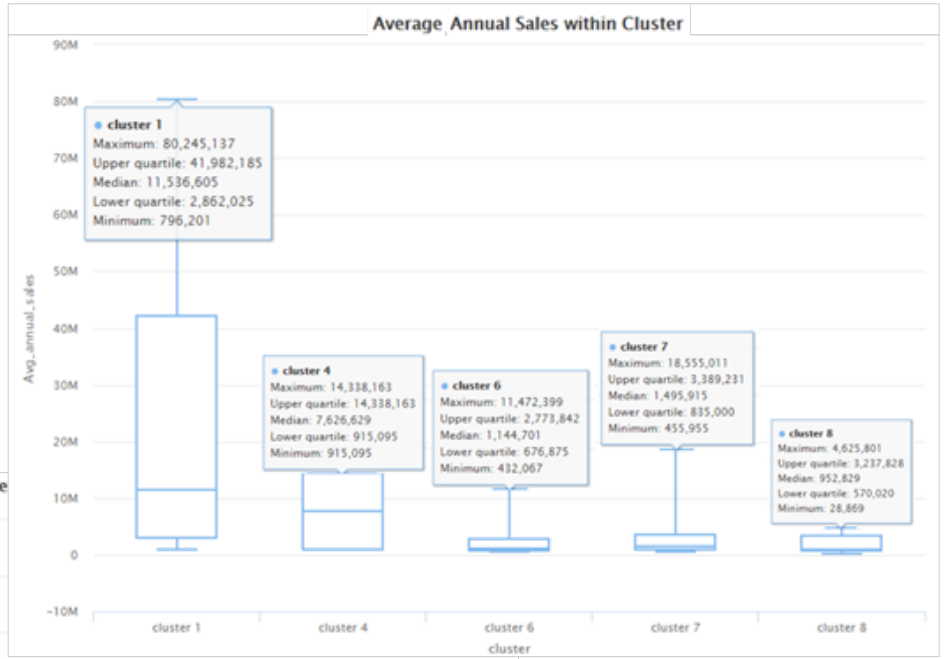


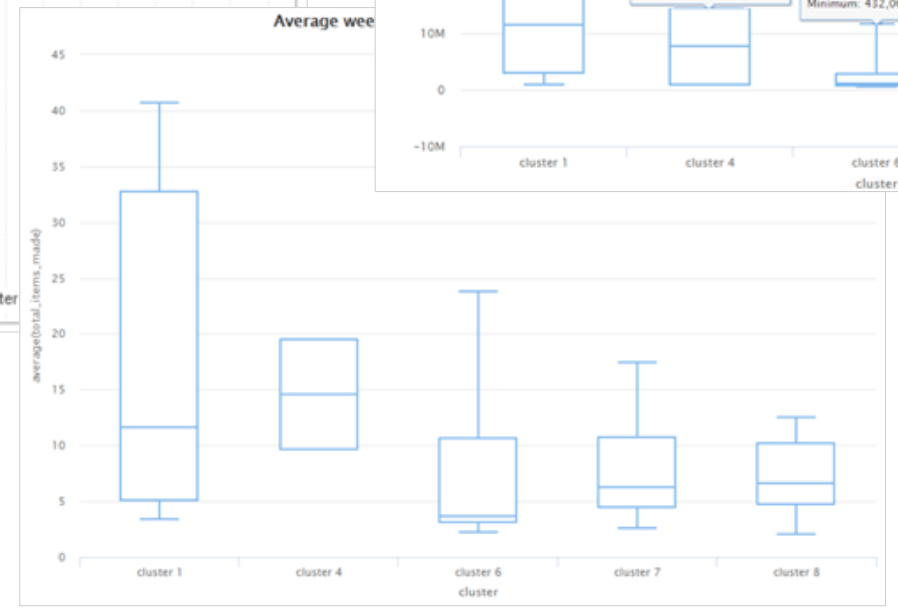
Centroid of cluster 7 and its behavioural patterns



Further Analysis







Behavioural similarity groups help understanding *market forces* and provide a convenient tool to characterise behaviours of direct competitors within each geographic area (left), their typical sales volume (middle), and a total value of sales within brand groups (right).

DEAKIN UNIVERSITY

Summary and References

- Time series shapelets are a novel tool for the analysis of time series data in science and engineering
- They can be useful for the analysis of business data
- Shapelets allow determination of behavioural patterns from sales data
- Shapelet clustering helps identifying market competitors responding to identical events
- Shapelet clusters can be explored to better understand the market and its forces

References

- Beggel, Laura, Bernhard X. Kausler, Martin Schiegg, Michael Pfeiffer, and Bernd Bischl. “Time Series Anomaly Detection Based on Shapelet Learning.” Computational Statistics 34, no. 3 (September 1, 2019): 945–76.
- Grabocka, J., Schilling, N., Wistuba, M., Schmidt-Thieme, L., 2014. Learning time-series shapelets. Proc. of the 20th ACM SIGKDD Int. Conf. on Knowledge discovery and data mining, ACM, pp. 392–401.
- Grabocka, J., Wistuba, M., Schmidt-Thieme, L., 2015. Scalable Discovery of Time-Series Shapelets. ArXiv:1503.03238.
- Kim, Sang, Hee Lee, Han Ko, Seung Jeong, Hyun Byun, and Kyong Oh. “Pattern Matching Trading System Based on the Dynamic Time Warping Algorithm.” Sustainability 10, no. 12 (2018): 4641.
- Whitfield, Troy W., Jie Wang, et al. “Functional Analysis of Transcription Factor Binding Sites in Human Promoters.” Genome Biology 13, no. 9 (September 5, 2012): R50.
- Zakaria, J., Mueen, A., Keogh, E., 2012. Clustering Time Series Using Unsupervised-Shapelets, in: 2012 IEEE 12th Int. Conf. on Data Mining, pp. 785–794.
- Zhao, Jiaping, and Laurent Itti. “ShapeDTW: Shape Dynamic Time Warping.” ArXiv:1606.01601 [Cs], June 5, 2016.

DEAKIN UNIVERSITY

Other Projects Involving IRI Data

- Comparison of deep learning and traditional methods of analysing multivariate time series
- Investigation of deep learning techniques to generate novel insights from multivariate sales and marketing data, which are not available from other techniques
- Investigation of machine learning methods suitable for wide rather than deep time-series data sets
- Application of lessons learnt to other types of financial data