

COMP3420: Advanced Databases and Data Mining

Mining data streams and time
series



Lecture outline

- Mining data streams

- Characteristics of data streams
- Stream data applications
- Data stream management system
- Challenges and methodologies of data stream processing
- Stream data mining versus stream querying
- Clustering data streams
- Challenges for mining dynamics in data streams

- Mining time-series data

- Categories of time-series movements
- Estimation of trend curve
- Trend discovery
- Similarity search in time-series analysis

Characteristics of data streams

- Data streams

- Continuous, ordered, changing, fast, huge amount
- Traditional DBMS—data stored in finite, persistent data sets

- Characteristics

- Huge volumes of continuous data, possibly infinite
- Fast changing and requires fast, real-time response
- Data stream captures nicely our data processing needs of today
- Random access is expensive—single scan algorithm (*can only have one look at each record!*)
- Store only the summary of the data seen thus far
- Most stream data are at pretty low-level or multi-dimensional in nature, needs multi-level (ML) and multi-dimensional (MD) processing

Stream data applications

- Telecommunication calling records
- Business: credit card transaction flows
- Network monitoring and traffic engineering
- Financial market: stock exchange
- Engineering & industrial processes: power supply and manufacturing
- Sensor, monitoring & surveillance: video streams, RFIDs
(Radio Frequency IDentification)
- Security monitoring
- Web logs and Web page click streams
- Massive data sets (even saved but random access is too expensive)

Architecture: Stream query processing

DSMS (Data Stream Management System)

Continuous query

Multiple streams

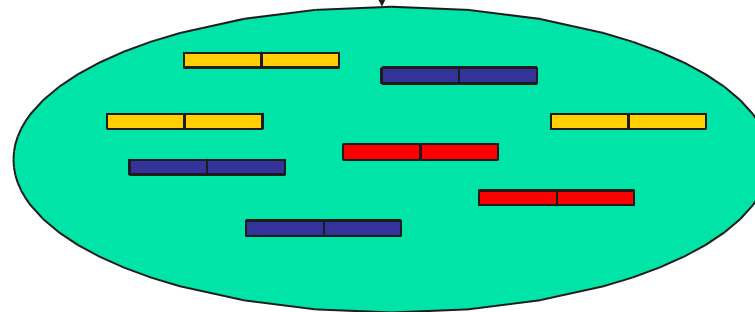


User/Application

Stream Query Processor

Results

Scratch Space
(Main memory and/or Disk)



DBMS versus DSMS

- Persistent relations
- One-time queries
- Random access
- “Unbounded” disk store
- Only current state matters
- No real-time services
- Relatively low update rate
- Data at any granularity
- Assume precise data
- Access plan determined by query processor, physical DB design
- Transient streams
- Continuous queries
- Sequential access
- Bounded main memory
- Historical data is important
- Real-time requirements
- Possibly multi-GB arrival rate
- Data at fine granularity
- Data stale/imprecise
- Unpredictable/variable data arrival and characteristics



Challenges of stream data processing

- Multiple, continuous, rapid, time-varying, ordered streams
- Main memory computations
- Queries are often continuous
 - Evaluated continuously as stream data arrives
 - Answer updated over time
- Queries are often complex
 - Beyond element-at-a-time processing
 - Beyond stream-at-a-time processing
 - Beyond relational queries (scientific, data mining, OLAP)
- Multi-level/multi-dimensional processing and data mining
 - Most stream data are at low-level or multi-dimensional in nature

Methodologies for stream data processing

- Major challenge
 - Keep track of a large universe (for example, IP address, not ages)
- Methodology
 - Synopses (trade-off between accuracy and storage)
 - Use *synopsis* data structure, much smaller ($O(\log^k N)$ space) than their base data set ($O(N)$ space), with N the number of elements in the stream data
 - Compute an *approximate answer* within a *small error range* (factor ϵ of the actual answer)
- Major methods
 - *Random sampling* (maintain a set of candidates in memory)
 - *Histograms* (approximate frequency distribution of values in stream)
 - *Sliding windows* (make decision based on only recent data)
 - *Multi-resolution models* (balanced trees, wavelets, micro-clusters)
 - *Sketches* (summarises data, can be done in one pass)
 - *Randomised algorithms* (Monte Carlo algorithm, bound on run time)



Stream data mining versus stream querying

- Stream mining is a more challenging task in many cases
 - It shares most of the difficulties with stream querying
 - But often requires less *precision*, for example, no join, grouping, sorting
 - Patterns are hidden and more general than querying
 - It may require exploratory analysis (not necessarily continuous queries)
 - Change in data characteristics: *Concept drift*
- Stream data mining tasks
 - Frequent patterns in data streams (approximate frequent patterns only)
 - Mining outliers and unusual patterns in stream data
 - Classification of stream data (approximate decision trees, classifier ensemble)
 - Clustering data streams



Clustering data stream methodologies

- Compute and store summaries of past data
- Apply divide-and-conquer strategy
 - Divide stream into chunks, summarise chunks, merge summaries
- Incremental clustering of incoming data (refine clusters)
- Perform micro-clustering as well as macro-clustering
 - Micro-clusters based on hierarchical bottom-up clustering
- Explore multiple time-granularity for cluster evolution
- Divide stream clustering into on- and off-line processes
 - Compute basic summaries of data snapshots online



Challenges for mining dynamics in data streams

- Most stream data is at a pretty low-level or multi-dimensional in nature: needs ML/MD processing
- Analysis requirements
 - Multi-dimensional trends and unusual patterns
 - Capturing important changes at multi-dimensions/levels
 - Fast, real-time detection and response
- Stream (data) cube or stream OLAP: Is this feasible?
 - Can we implement it efficiently?
 - More details in text book

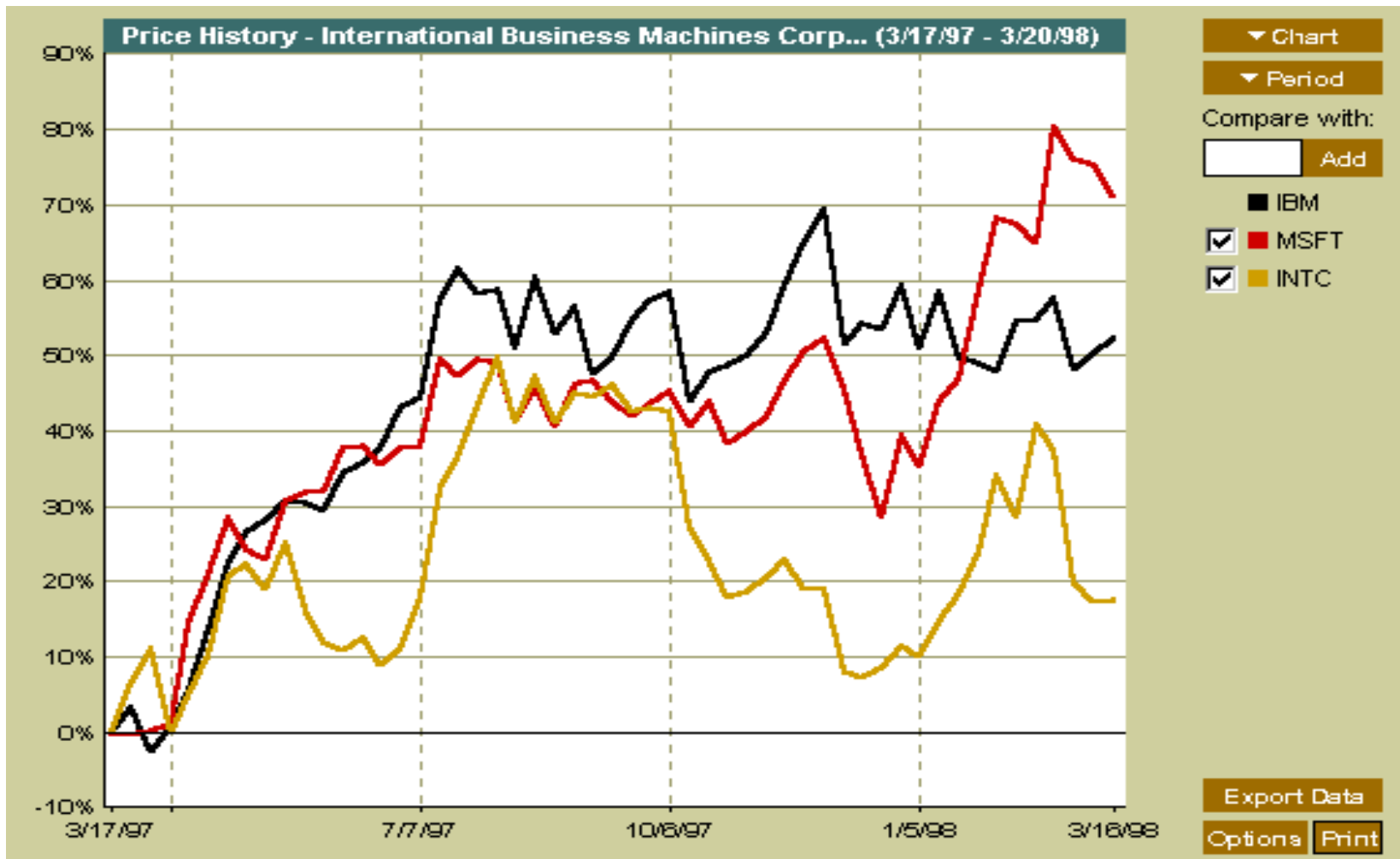


Multi-dimensional stream analysis: Examples

- Analysis of Web click streams
 - Raw data at low levels: seconds, Web page addresses, user IP addresses, IP port numbers, ...
 - Analysts want: changes, trends, unusual patterns, at reasonable levels of details
 - For example: *Average clicking traffic in North America on sports in the last 15 minutes is 40% higher than that in the last 24 hours*
- Analysis of power consumption streams
 - Raw data: power consumption flow for every household, every minute
 - Patterns one may find: *average hourly power consumption surges up 30% for manufacturing companies in Chicago in the last 2 hours today than that of the same day a week ago*

Mining time-series data

- Time-series database
 - Consists of sequences of values or events changing with time
 - Data is recorded at *regular intervals*
 - Characteristic time-series components: Trend, cycle, seasonal, irregular
- Applications
 - Financial: stock price, inflation
 - Industry: power consumption
 - Scientific: experiment results
 - Meteorological: precipitation
 - Medical treatments



- A time series can be illustrated as a time-series graph which describes a point moving with the passage of time

Categories of time-series movements

- Long-term or trend movements (trend curve) (T)
 - General direction in which a time-series is moving over a long interval of time
- Cyclic movements or cycle variations (C)
 - Long term oscillations about a trend line or curve
 - For example, business cycles, may or may not be periodic
- Seasonal movements or seasonal variations (S)
 - Almost identical patterns that a time series appears to follow during corresponding months of successive years
- Irregular or random movements (I)
- Time series analysis: decomposition of a time-series into these four basic movements
 - *Additive Model*: $TS = T + C + S + I$
 - *Multiplicative Model*: $TS = T \times C \times S \times I$

Estimation of trend curve

- The freehand method
 - Fit the curve by looking at the graph
 - Costly and barely reliable for large-scaled data mining
- The least-square method
 - Find the curve minimising the sum of the squares of the deviation of points on the curve from the corresponding data points
- The moving-average method
 - Smooths the data
 - Eliminates cyclic, seasonal and irregular movements
 - Loses the data at the beginning or end of a series
 - Sensitive to outliers (can be reduced by weighted moving average)

$$\frac{y_1 + y_2 + \dots + y_n}{n}, \frac{y_2 + y_3 + \dots + y_{n+1}}{n}, \frac{y_3 + y_4 + \dots + y_{n+2}}{n}, \dots$$



Trend discovery: Estimation of seasonal variations

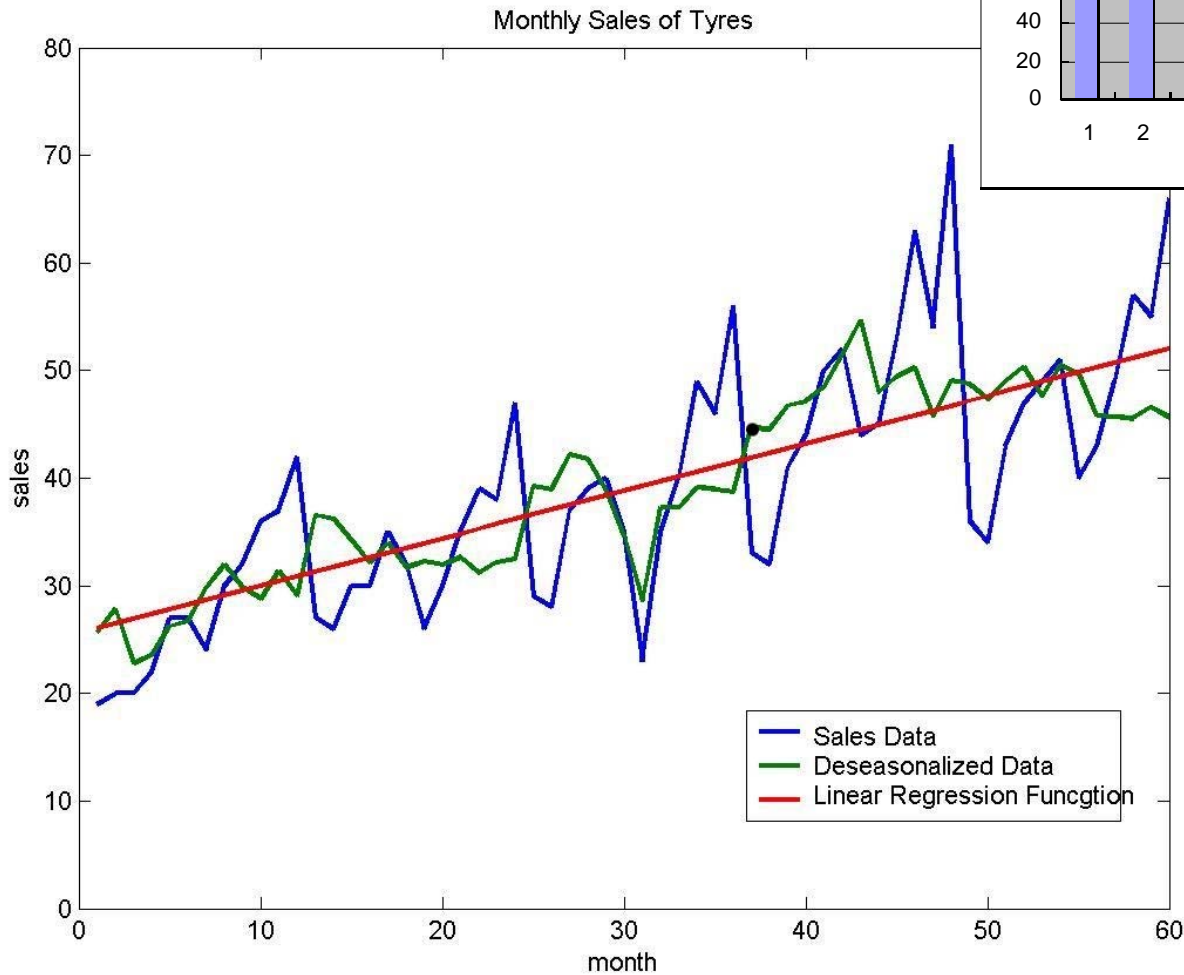
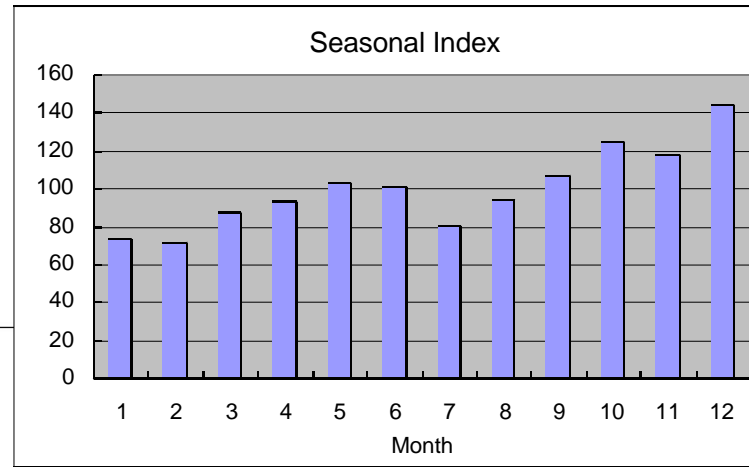
- **Seasonal index**

- Set of numbers showing the relative values of a variable during the months of the year
- For example, if the sales during October, November, and December are 80%, 120%, and 140% of the average monthly sales for the whole year, respectively, then 80, 120, and 140 are seasonal index numbers for these months

- **De-seasonalised data**

- Data adjusted for seasonal variations for better trend and cyclic analysis
- Divide the original monthly data by the seasonal index numbers for the corresponding months

Seasonal index example



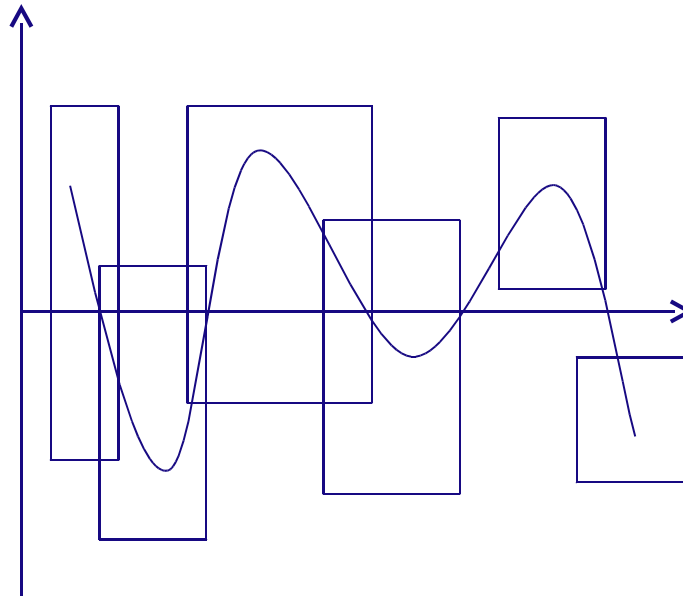
Raw data from:
http://www.bbk.ac.uk/mano/p/man/docs/QII_2_2003%20Time%20series.pdf

Trend discovery (2)

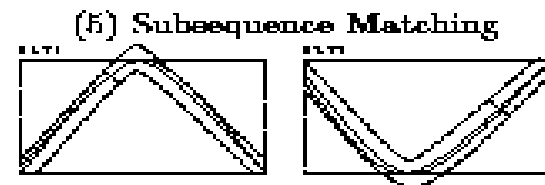
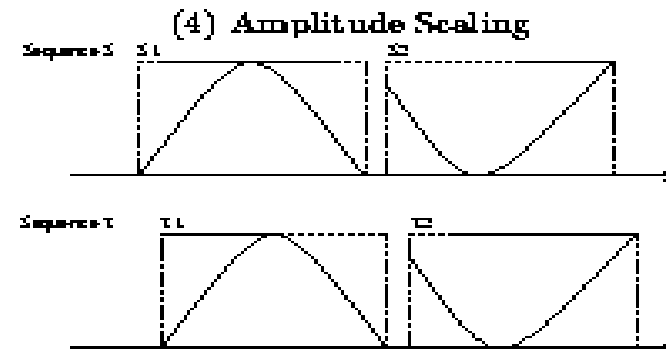
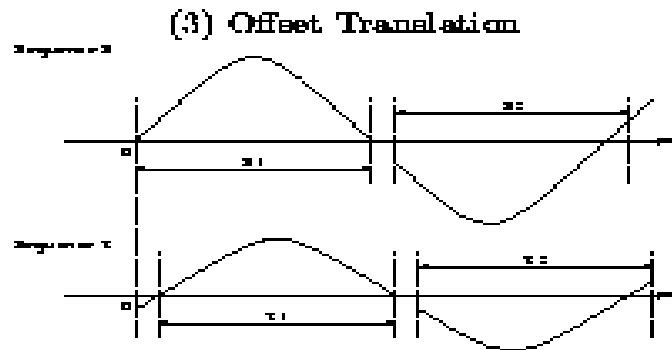
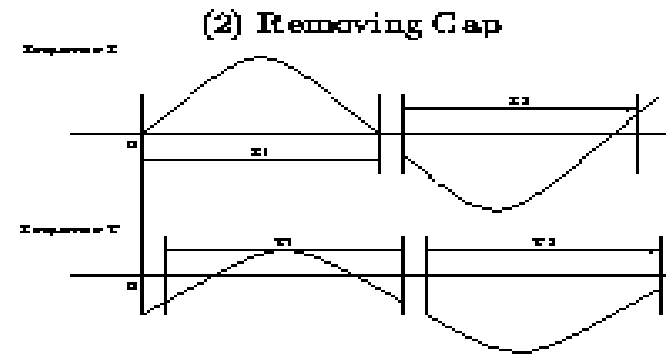
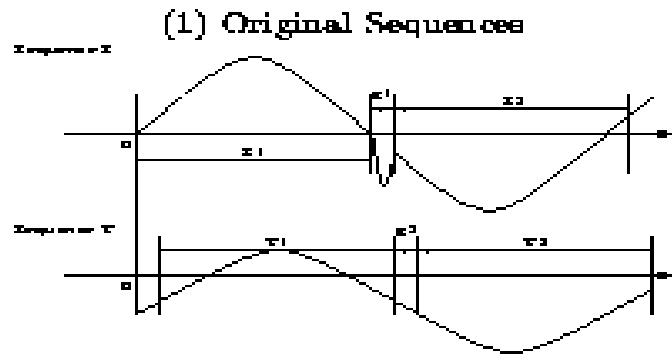
- Estimation of cyclic variations
 - If (approximate) periodicity of cycles occurs, cyclic index can be constructed in much the same manner as seasonal indices
- Estimation of irregular variations
 - By adjusting the data for trend, seasonal and cyclic variations
- With the systematic analysis of the trend, cyclic, seasonal, and irregular components, it is possible to make long- or short-term predictions with reasonable quality

Similarity search in time-series analysis

- Normal database query finds exact match
- Similarity search finds data sequences that differ only slightly from the given query sequence
- Two categories of similarity queries
 - *Whole matching*: find a sequence that is similar to the query sequence
 - *Subsequence matching*: find all pairs of similar sequences
- Typical Applications
 - Financial market (similar stocks)
 - Market basket data analysis
 - Scientific databases
 - Medical diagnosis



Analysis of similar time series

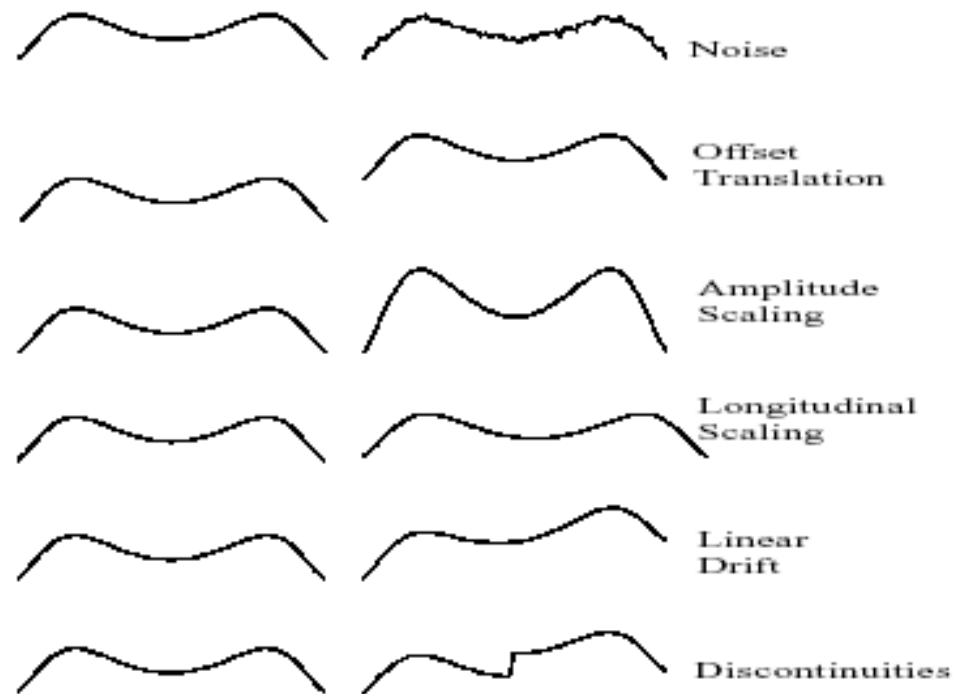


Case study: Stationary Stores

- Denny and Lee, V. C. S. (2004). An alternative methodology for mining seasonal pattern using self-organizing map. PAKDD.
 - Real-life data from a major stationery retail chain
 - Two year sales data, 1999 and 2000, from seven branches
 - About 1.5 millions of transactional data for 17,836 products
 - 34 groups of similar items (such as school equipments), and further divided into 551 sub-groups of more similar items (such as pencil cases).

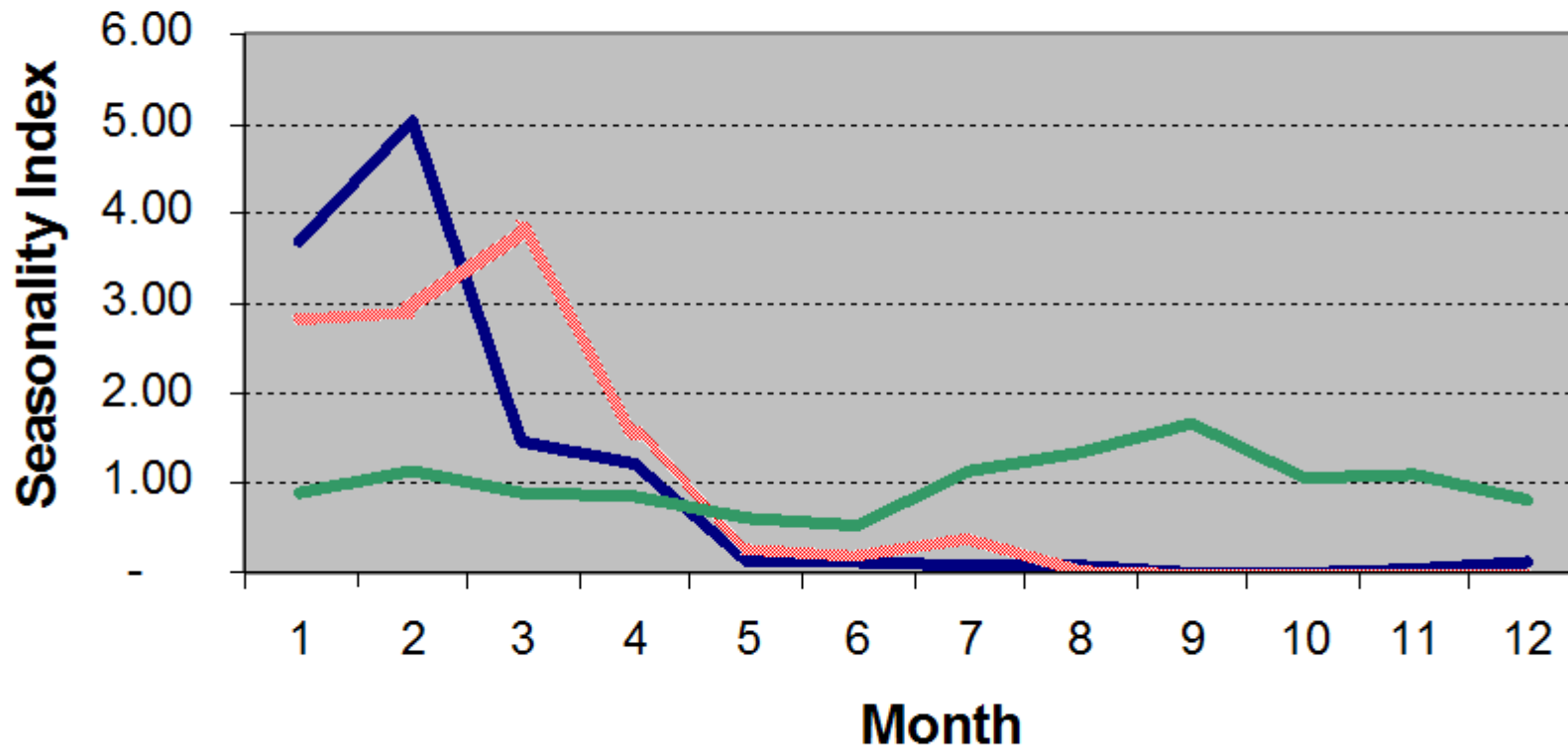
Case study: difficulties in comparing time series

- Longitudinal scaling and linear drift can be considered not important in comparing seasonality pattern.
- Discontinuities and noise can be alleviated by aggregating daily sales quantity in a period of time
- Amplitude scaling and offset translation problems can be solved by performing normalization of the data set.

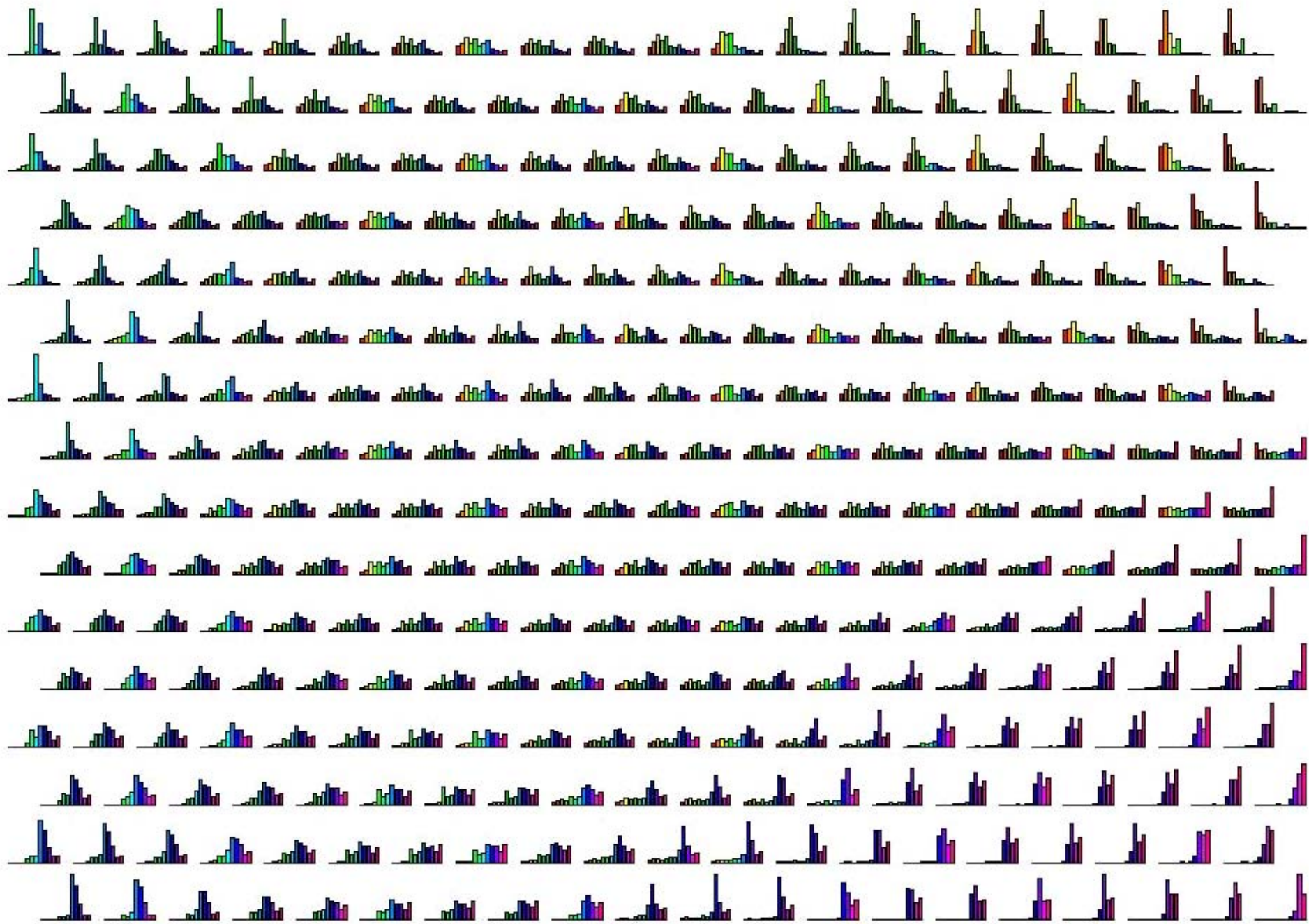


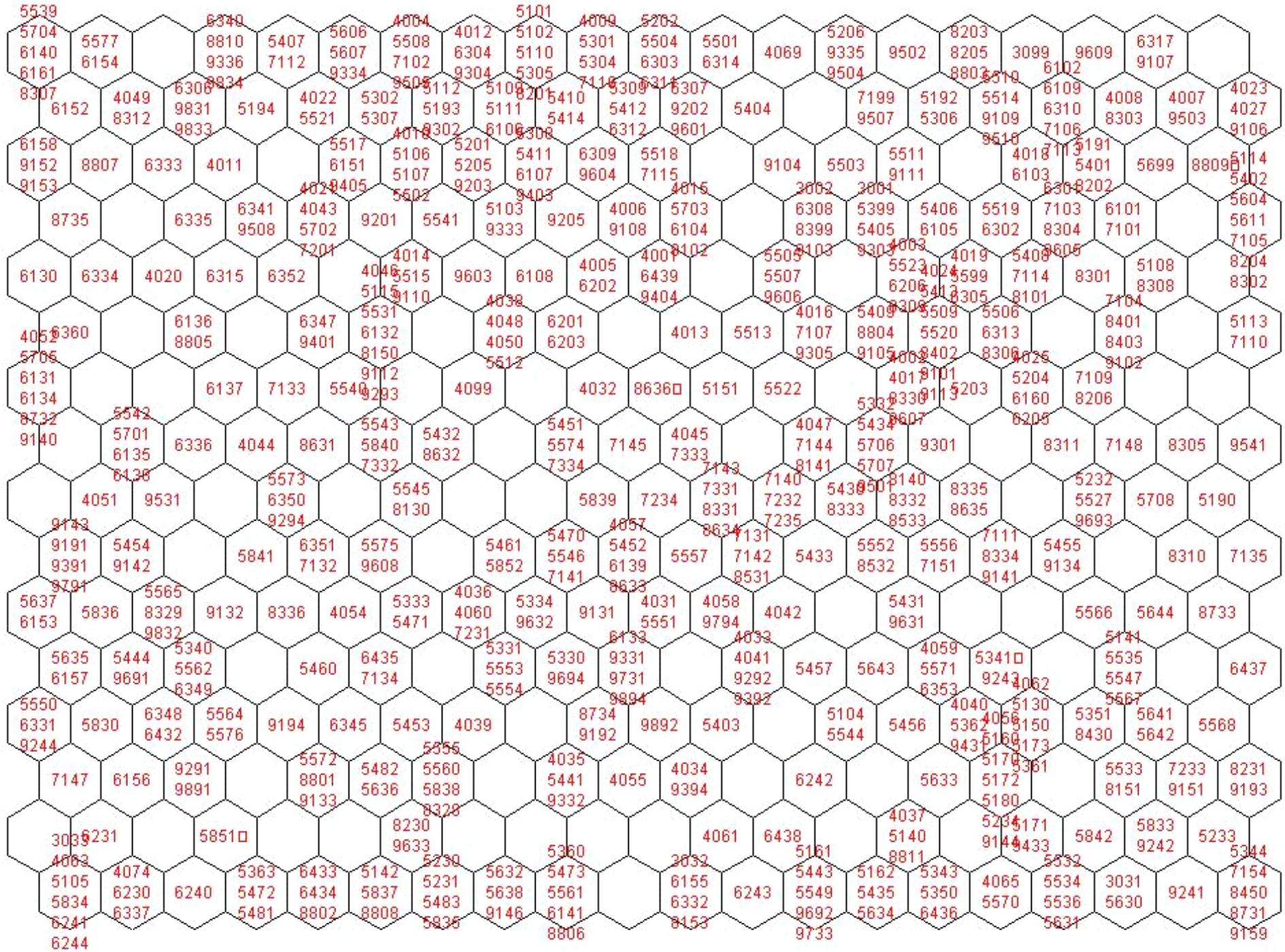
Case Study: Data Preprocessing

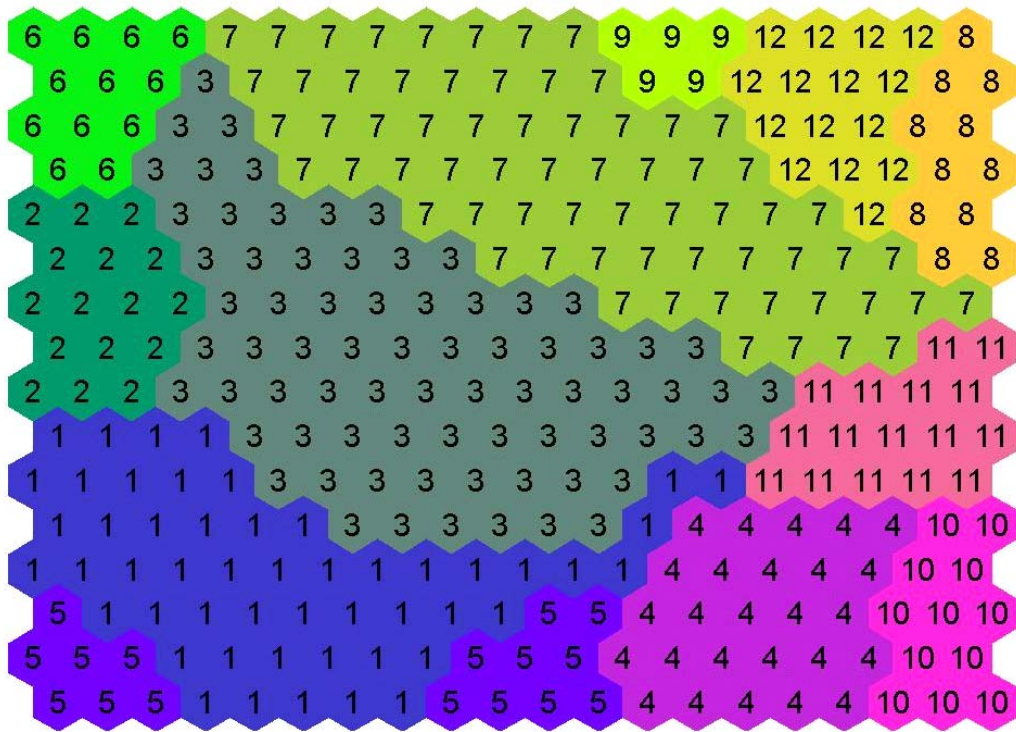
Code	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Avg.
4007	2,814	3,843	1,106	939	89	86	70	49	15	8	23	104	762
6109	477	487	639	256	45	31	70	5	1	0	0	0	168
4031	392	481	387	375	262	220	481	584	711	459	475	348	431



— 4007 - Pencil — 6109 - Writing Book in pack — 4031 - Technic Pens







No	Total Member	Total Sales Qty in 2 Years			Average of Seasonal Index											
		Average	Sum	%	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1	79	11,862	937,126	12.2%	0.1	0.1	0.1	0.2	0.7	1.2	1.6	1.5	1.9	1.7	1.6	1.2
2	20	22,953	459,055	6.0%	0.2	0.3	0.3	0.4	0.5	1.1	4.8	1.7	0.9	0.7	0.6	0.4
3	127	21,668	2,751,805	35.9%	0.6	0.7	0.9	1.0	0.9	1.0	1.2	1.3	1.2	1.1	1.0	0.9
4	56	3,318	185,820	2.4%	0.0	0.0	0.0	0.0	0.1	0.1	0.2	0.2	2.4	3.8	2.9	2.4
5	25	3,223	80,583	1.1%	0.1	0.1	0.1	0.1	0.2	0.5	0.7	2.2	4.0	1.9	1.3	0.8
6	23	4,732	108,834	1.4%	0.1	0.1	0.1	0.5	1.6	4.0	1.9	1.5	0.8	0.6	0.4	0.3
7	145	17,467	2,532,653	33.0%	1.3	1.4	1.5	1.4	1.0	0.9	1.1	0.9	0.7	0.6	0.6	0.5
8	18	3,165	56,968	0.7%	6.4	2.5	0.8	0.6	0.5	0.2	0.2	0.3	0.3	0.1	0.1	0.1
9	7	5,120	35,839	0.5%	0.2	1.2	2.7	5.4	0.8	0.4	0.6	0.1	0.2	0.2	0.1	0.1
10	21	2,107	44,242	0.6%	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	2.3	5.7	3.2
11	17	16,824	286,005	3.7%	0.9	0.7	0.5	0.5	0.5	0.8	1.0	0.7	1.0	1.0	1.8	2.7
12	34	5,565	189,200	2.5%	2.1	3.3	2.7	1.5	0.8	0.4	0.4	0.3	0.2	0.2	0.1	0.1

