Automated pattern mining with a scale dimension

Jan M. Żytkow† & Robert Zembowicz

Computer Science Department, Wichita State University, Wichita, KS 67260-0083
† also Institute of Computer Science, Polish Academy of Sciences, Warsaw
{zytkow, robert}@cs.twsu.edu

Abstract

An important but neglected aspect of automated data mining is discovering patterns at different scale in the same data. Scale plays the role analogous to error. It can be used to focus the search for patterns on differences that exceed the given scale and to disregard those smaller. We introduce a discovery mechanism that applies to bi-variate data. It combines search for maxima and minima with search for regularities in the form of equations. Groups of detected patterns are recursively searched for patterns on their parameters. If the mechanism cannot find a regularity for all data, it uses patterns discovered from data to divide data into subsets, and explores recursively each subset. Detected patterns are subtracted from data and the search continues in the residua. Our mechanism seeks patterns at each scale. Applied at many scales and to many data sets, it seems explosive, but it terminates surprisingly fast because of data reduction and the requirements of pattern stability. We walk through an application on a half million datapoints, showing how our method leads to the discovery of many extrema, equations on their parameters, and equations that hold in subsets of data or in residua. Then we analyze the clues provide by the discovered regularities about phenomena in the environment in which the data have been gathered.

Automated data mining: the role of error and scale

Regularities at different scale (also called tolerance) are common in data, when large amounts of datapoints are available. For instance, consider a time series in which the overall linear growth may be altered with a short cycle periodic pattern. Both patterns may be caused by different phenomena. It is possible that those phenomena can be recognized from the discovered patterns. Real data contain information about many phenomena as a rule, not as an exception.

Statistical data analysis offers many methods for data exploration that assist human data miners (Tukey 1977; Hoaglin, Mosteller & Tukey 1983), yet the majority of methods make a small step and require user choice of the next step (Kendall & Ord, 1990). Today's statistical packages, such as Lisp-Stat (Tierney, 1990) offer the user access to medium level operators, but only recently, in the domain of Knowledge Discovery, the research focused on large-scale search for knowledge that can be invoked with a simple command and keeps the user out of the loop (Piatetsky-Matheus, 1991; Kloesgen 1996; 49er: Zytkow Zembowicz, 1993). Such systems become necessary when thousands of exploratory steps are needed.

The need for mining massive data at many scale levels leads to a challenging vision for automated knowledge discovery: develop a mechanism that discovers as many independent patterns as possible, not overlooking patterns at any scale but not accepting spurious regularities which can occur by chance when data are confronted with very large hypotheses spaces. That mechanism should be able to detect patterns of different types, such as equations, maxima and minima, and search for patterns in data subsets if the search fails to detect patterns satisfied by all data.

In this paper we present an automated data mining mechanism that works efficiently for large amounts of data and makes progress on each of these requirements. We report an application of our methods on a large set of bi-variate data, computationally efficient and leading to a number of surprising conclusions. We concentrate on the algorithmic aspect of our system. Because of the paper size limit we do not present the representation of the nascent knowledge and the way in which the unknown elements of knowledge drive the search. This mechanism is a modification of FAHRENHEIT's knowledge representation (Żytkow, 1996).

The roles of error in pattern discovery

Error is a parameter that defines the difference between values of a variable which are empirically non-distinguishable. Scale plays the role analogous to error. It specifies the difference between the values of a variable which we, tentatively, consider unimportant. Because of this analogy, in order to search for patterns at different scale we can simply replace the error with scale in our discovery algorithms.

Let us briefly summarize the roles of error (noise) in the process of automated discovery. Those roles are well-known in statistical data analysis. Consider the data in the form $D = \{(x_i, y_i, e_i) : i = 1, ..., N\}$. Consider the search for equations of the form $y = f(x)$ that fit data (x_i, y_i) within the accuracy of error e_i. For the same $\{(x_i, y_i) : i = 1, ..., N\}$, the smaller are the error values $e_i, i = 1, ..., N$, the closer fit is required between data and equations. Even a constant pattern $y = C$ can fit any data, if error is very large, but the smaller is the error, the more complex equations may be needed to fit the data.

The same conclusions about error apply to other patterns, too. Let us consider maxima and minima. For the same data (x_i, y_i), when the error is large, many differences between the values of y are treated as random fluctuations within error. In consequence, few maxima and minima are detected. When the error is small, however, the same differences may become significant. The smaller is the error, the larger number of maxima and minima shall be detected.

Knowledge of error has been used in many ways during the search conducted by Equation Finder (**EF**: Zembowicz & Żytkow, 1992):

1. The error is used in the evaluation of each equation $y = f(x)$. For each datum, $0.5e_i = \sigma_i$ can be interpreted as the standard deviation of the normal distribution $N(y(x_i), \sigma_i)$ of y's for each x_i, where $y(x_i)$ is the mean value and σ_i is standard deviation. Knowledge of that distribution permits to compute the probability that the data have been generated as a sum of the value $y(x_i)$ and the value drawn randomly from the normal distribution $N(0, \sigma_i)$, for all $i = 1, \ldots, N$.

2. When the error e_i varies for different data, the weighted χ^2 value $[(y_i - f(x_i))/0.5e_i]^2$ is used to compute the best fit parameters for any model. This enforces better fit to the more precise data.

3. Error is propagated to the parameter values of $f(x)$. For instance, EF computes the error e_a for the slope a in $y = ax$. When patterns are sought for those parameters (e.g., in BACON-like recursive search for multidimensional regularities), the parameter error is used as data error.

4. If the parameter error is larger than the value of that parameter, EF assumes that the parameter value is zero. When $y = ax + b$ and $|a| < e_a$, then the equation is reduced to a constant.

5. EF generates new variables by transforming the initial variables x and y into terms such as $\log(x)$. Error values are propagated to the transformed data and used in search for equations that apply those terms.

Phenomena at different scale

When we collect data about a physical, social or economical process, it is common that the data capture phenomena at different scale. They describe the process, but the values may be influenced by data collection, instruments behavior, and the environment. Different phenomena can be characterized by their scales in both variables. Periodic phenomena, for instance, can have different amplitudes of change in y and different cycle length in x. In a time series (x is time) one phenomenon may occur in a daily cycle, the cycle for another can be few hours, while still another may follow a monotonous dependence between x and y. Each phenomenon may produce influence of different scale on the value of y.

Each datum combines the effects of all phenomena. When they are additive, the measured value of y for each x is the total of values contributed by each phenomenon. Given the resultant data, we want to separate the phenomena by detecting patterns that describe each of them individually. The basic question of this paper is how can it be done by an automated system.

Suppose that a particular search method has captured a pattern P in data D. Subtracting P from D produces residua which hold the remaining patterns. Repeated application of pattern detection and subtraction can gradually recover patterns that account for several phenomena. In this process, one has to be cautious of artifacts generated by spurious patterns and propagated to their residua.

It is a good idea to start the search for patterns from large scale phenomena, by using a large value of scale in pattern finding. The phenomena captured at the large values of error follow simple patterns. Many smaller scale patterns can be discovered later, in the residua obtained by subtracting the larger scale patterns.

The roles for maxima and minima

In this paper we concentrate on two types of patterns: maxima/minima, jointly called extrema, and equations. We explore the ways in which the results in one category can feedback the search for patterns of the other type.

A simple algorithm can detect maxima and minima at a given scale δ:

Algorithm: Find Extrema (X_{max}, Y_{max}) **and** (X_{min}, Y_{min})

```
given ordered sequence of points (x_i, y_i), i = 1 . . N, and scale δ
   task ← unknown, X_max ← x_1, X_min ← x_1, Y_max ← y_1, Y_min ← y_1
   for i from 2 to N do
      if task ≠ max and y_i > Y_min + δ then
         store minimum (X_min, Y_min)
         task ← max, X_max ← x_i, Y_max ← y_i
      else if task ≠ min and y_i < Y_max − δ then
         store maximum (X_max, Y_max)
         task ← min, X_min ← x_i, Y_min ← y_i
      else if y_i > Y_max then X_max ← x_i, Y_max ← y_i
      else if y_i < Y_min then X_min ← x_i, Y_min ← y_i
   if Y_max − Y_min > δ then ; handle the last extremum, if any
      if task = min then store minimum (X_min, Y_min)
      else if task = max then store maximum (X_max, Y_max)
```

Our discovery mechanism uses maxima and minima detected by this algorithm in several ways. First, if the number of extrema is M, then the minimum polynomial degree tried by Equation Finder is $M + 1$. A degree higher than M may be ultimately needed, because the inflection points also increase the degree of the polynomial. As the high degree polynomials are difficult to interpret and generalize, if $M > 3$, only the periodic functions are tried.

Another application is search for regularities on different properties of maxima and minima, such as the location, and height. Those regularities can be instrumental in understanding the nature of a periodic phenomenon. A regularity on the locations of the subsequent maxima and minima estimates the cycle length. A regularity on the extrema heights estimates the amplitude of a periodic pattern. Jointly, they can guide the search for sophisticated periodic equations.

Still another application is data partitioning at the extrema locations. Data between the adjacent extrema, detected at scale δ, are monotonous at δ, so that the equation finding search similar to BACON1 (Langley et al., 1987) applies in each partition. We use EF, limiting the search to linear equations and term transformations to monotonous (e.g., $x' = \log x, x' = \exp x$). Since all the data in each partition fit a constant at the tolerance level $1/2 \times \delta$, EF is applied with the error set at $1/3 \times \delta$.

Data reduction

The search for extrema at all scales starts at the minimum positive distance in y between adjacent datapoints, or the error of y, if it is known. The search uses the minimum difference between the adjacent extrema as the next scale value. The search terminates at the first scale at which no extrema have been found. Since an extremum at a larger scale must be also an extremum at each lower scale, when the search proceeds from the low end of scale, the extrema detected at a given scale become the input data for the search at the next higher scale. This way the number of data is reduced very fast and the search at all levels can be very efficient. The whole search for extrema at each scale typically takes less than double the time spent at the initial scale.

Algorithm: Detect extrema at all tolerance levels

Given an ordered sequence of points (x_i, y_i), $i = 1 \ldots N$,
  $\delta_{min} \leftarrow \left|\min(y_i - y_{i+1})\right|$, for all $y_i \neq y_{i+1}$, $i = 1 \ldots N$
  $\delta \leftarrow \delta_{min}$, data $\leftarrow$ (x_i, y_i), $i = 1 \ldots N$
  while data includes more than two points **do**
    Find Extrema $(X_{min/max}, Y_{min/max})$ **in data at scale** δ
    Store δ , store list-of-extrema $(X_{min/max}, Y_{min/max})$
    $\delta \leftarrow \delta + \delta_{min}$, data $\leftarrow$ list-of-extrema $(X_{min/max}, Y_{min/max})$

The search for equations may benefit from another type of data reduction. A number of adjacent data, at a small distance between their x values, can be binned together and represented by their mean value and standard deviation. The size of the bin depends on the tolerance level in the x dimension. The results of binning are visualized in Figures 1 and 2. Each pixel in the x dimension summarizes about 500 data, so that about 0.5 mln data have been reduced to about 1000.

Pattern stability

No scale is a priori more important than another, as patterns can occur at any scale. When the search for patterns is successful at scale δ, the same pattern can often be detected at scales close to δ. Patterns that hold at many levels are called stable (Witkin, 1983). Consider the stability of extrema. Each extremum that occurs at a given level must also occur at all lower levels. Stability applies to pairs of adjacent extrema and is measured by the range of scale levels over which a given min/max pair is detected as adjacent. Expanding the definition to the set of all extrema at a given scale, we measure stability as the range of scale levels over which the set of extrema does not change.

Stability is important for several reasons. (1) As discussed earlier, when a number of extrema are detected at a given scale, regularities can be sought on their location, height, and width. It is wasteful to detect the same regularities many times at different scales and then realize that they are identical. A better idea is to recognize that the set of extrema is stable over an interval of scale levels, and search for regularities only once. (2) A stable pattern is a likely manifestation of a real phenomenon, while an unstable set of extrema S may be an artifact. Some extrema may be included in S by chance. A slight variation in the tolerance level removes them from S. The search for regularities in such a set S may fail or lead to spurious regularities. It would be a further waste of time to seek their interpretation and generalization.

Since regularities for extrema may lead to important conclusions about the underlying phenomena, our system pays attention to sets of extrema which are stable across many tolerance levels. It searches the sets of extrema and picks the first stable set at the high end of scale.

Algorithm: Detect stable set of extrema at high end of scale

Given a sequence of extrema sets EXTREMA$_i$, ordered by δ_i,
$\delta_{min} \leq \delta_i \leq \delta_{max}$:
  for δ **from** δ_{min} **to** δ_{max} **do**
    Compute the number E_i of extrema in EXTREMA$_i$
  for each different E_i **do**
    Compute the number N_i of occurrences of E_i ;; The higher
    ;; is N_i, the more stable the corresponding set EXTREMA$_i$
  Let N_a, N_b two highest numbers among N_i
  return EXTREMA$_i$ for the minimum(E_a, E_b)
    ;; That among two most stable sets which is of higher scale
    and return δ_{stable} = average(δ) for EXTREMA$_i$

Residual data

Our mechanism treats patterns as additive. It subtracts the detected patterns from data and seek further patterns in the residua. We will now present the details of subtraction and discuss termination of the search.

Both equations and extrema are functional relationships between x and y: $y = f(x)$. Some extrema can be described by equations that cover also many data that extend far beyond a given extremum. For instance, a second degree polynomial that fits one extremum, may at the same time capture a range of data. Equations may not be found, however, for many extrema. Those we represent point by point. In the first case, when $y = f(x)$ is the best and acceptable equation, the residua are computed as $r_i = y_i - f(x_i)$, and they oscillate around $y = 0$. In the second case we remove from the data all datapoints that represent the extremum.

In the first case the data are decomposed into pattern $y = f(x)$ and residua $\{(x_i, r_i), i = 1, \ldots, N\}$, so that $y_i = f(x_i) + r_i$. In the second case, the data are partitioned into the extremum $\{(x_i, y_i) : s(x_i)\}$, where $s(x)$ describes the scope of the extremum, and the residua $\{(x_i, y_i) : \neg s(x_i)\}$.

If the residua (x_i, r_i) deviate from the normal distribution, the search for patterns applies recursively. Eventually no more patterns can be found in residua. This may happen because regularities in the residua are not within the scope of search or because the residua represent Gaussian noise. In the latter case, the final data model is $y = f(x) + N(0, \varepsilon(x)$, where $\varepsilon(x)$ is Gaussian, while $f(x)$ represents all the detected patterns. Since the variability of residua is much smaller than that of the original data, the subtraction of patterns typically takes only a few iterations.

```
Algorithm: Detect Patterns
  D ← the initial data
  until data D are random
    seek stable pattern(s) P in D at the highest tolerance levels
    D ← subtract pattern(s) P from D
```

Data

As an example we will consider a large number of data collected in a simple experiment. In order to find the theory of measurement error for an electronic balance, we automatically collected the mass readings of an empty beaker placed on the balance. The measurements have been continued for several days, approximately one datapoint per second. Altogether we got 482,450 readings (7.2 megabytes). Fig. 1 illustrates the data. The beaker weighted about 249.130mg. Since the nominal accuracy of the balance is 0.001g=1mg, we expected a constant reading that would occasionally diverge by a few milligrams from the average value. But the results were very surprising. All datapoints have been plotted, but many adjacent data have been plotted at the same value of x.

We can see a periodic pattern consisting of several large maxima with a slower ascent and a more rapid descent. The heights of the maxima seem constant or slightly growing, and they seem to follow a constant cycle. Superimposed on this constant pattern are smaller extrema of different height. Several levels of even smaller extrema are not visible in Figure 1, because the time dimension has been compresses, but they are clear in the original data.

Upon closer examination, one can notice seven data points at the values about 249.430g, that is 0.3g above the average result, a huge distance in comparison to the accuracy of the balance. They can be seen as small plus signs in the upper part of Figure 1. There is one point about 0.3g below the average data. Those data must have been caused by the same phenomenon because they are very special and very similar: they are momentary peaks of one-second duration, at the similar distance from the bulk of the data.

Apparently, several phenomena must have contributed to those data. Can all these patterns be detected in automated way by a general purpose mechanism, or are our eyes smarter than our computer programs? Can our mechanism find enough clues about those patterns to discover the underlying phenomena?

Results of search for patterns at many scales

We will now illustrate the application of our multi-scale search mechanism on the dataset described in the previous section.

Detect extrema at all tolerance levels: The search for extrema at all levels of tolerance started from 482,450 data and iterated through some 300 scale levels. Since the number of extrema decreased rapidly between the low scale levels, the majority of time has been spent on the first pass through the data. The number of extrema at the initial scale of $1mg$ has been 29,000, so the data have been reduced to 6%. The number of extrema has been under 50 for $\delta > 9mg$.

Find the first stable set of extrema: The algorithm that finds the first stable set has been applied to all 300 sets of extrema. It detected a set stable at the tolerance levels between 30 and 300. The number of extrema has been 17, including 8 maxima and 9 minima. Eight of those are the outliers discussed in Section 2, while the complementary extrema lie within the main body of data. The stable extrema became the focus of the next step. We will focus on the maxima, where the search has been successful.

Use the stable extrema set: A simple mechanism for identification of similar patterns (Żytkow, 1996) excluded one maximum, which differed very significantly from all others in both the height and width (to limit the size of this paper, we do not discuss the extrema widths). That maximum is an artifact accompanying the minimum located $300mg$ under the bulk of the data. When applied to the remaining maxima, the equation finder discovered two strong regularities: (1) maxima heights are constant, (2) the maxima widths are constant (equal one second). No equation has been found for their location. Even if these 1-second deviations from the far more stable readings of mass occurred only seven times in nearly half million data, the pattern they follow may help us sometime to identify their cause.

Subtract patterns from data: Since the seven maxima have the width of 1 datapoint each, according to section 1.6 they have been removed from the original data. 482,442 data remained for further analysis (one single-point minimum has been removed, too).

Detect patterns in the residua: The search for patterns continued recursively in the residual data. Now the extrema have been found at the much more limited range of tolerance levels, between 1 and 43. The numbers of maxima at each scale have been depicted in Table 1 in the rows labeled M_{max}. The stability analysis determined a set of five maxima and the corresponding minima, which have been stable at the scale between 23 and 30. No regularity has been found for extrema locations and amplitudes, but interesting reg-

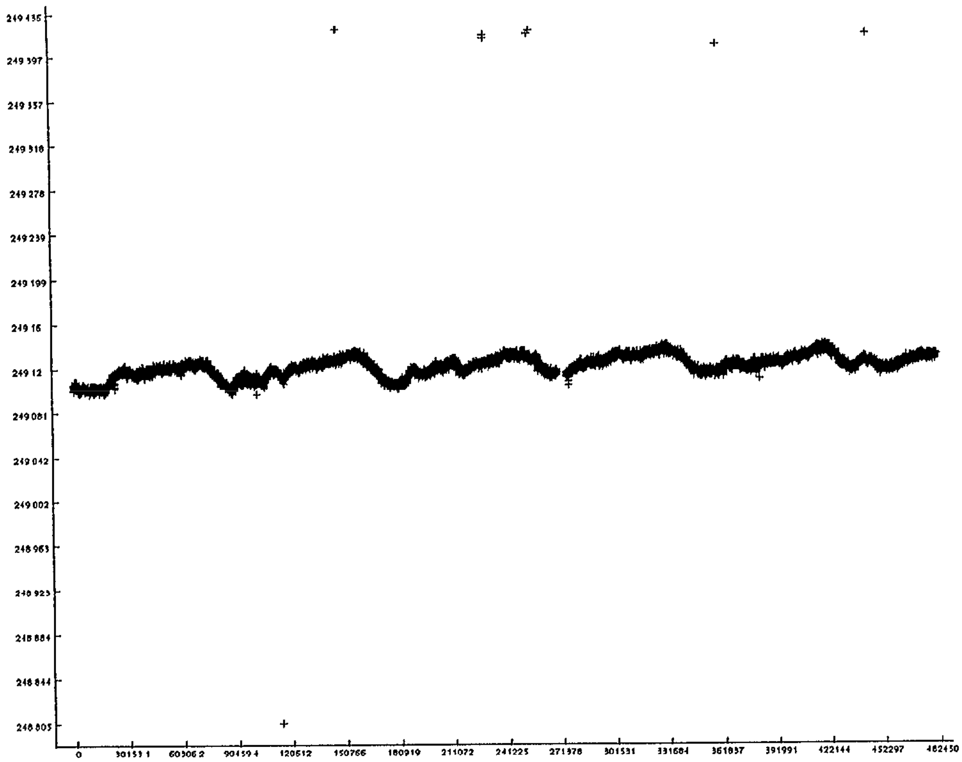

Figure 1: The raw results of mass measurements on an electronic balance over the period of five days (482450 readings).

ularities have been found for heights and locations of the maxima, from the data listed in Table 2.

Max. number	1	2	3	4	5
Height	249.137	249.144	249.143	249.148	249.148
Location	69944	155970	239367	328073	418167

Table 2: The first stable group of maxima, at the scale 23–30 in Table 1.

For those five maxima, stable equations have been found for location and height as functions of the maxima number, M: *location* $= 86855M - 18261$ and ***height*** $= 0.0026M + 249.144$.

The successful regularities have been also found for minima locations and heights. But no equation has been found for all data at the tolerance level $1/6 \times 30mg = 5mg$.

Partition the data: Since no trigonometric equation has been found (cf. section 1.3), the data have been partitioned at the extrema, and search for monotonous equations has been tried and succeeded in each partition, with different equations. For instance, in two segments of data: down-1 (in Figure 2, data for X from about 70,000 to about 90,000) and down-2 (X from 160,000 to 180,000), the equations are linear:

(segment down-1) $y = -1.61 \times 10^{-6} \times x + 249.249$
(segment down-2) $y = -1.46 \times 10^{-6} \times x + 249.371$

By subtracting these regularities from the data, two sets of residua have been generated, labeled down-1 and down-2. Further search, applied to residua in each partition, revealed extrema at lower tolerance levels. The patterns for the stable maxima locations as a function of maxima number M, for instance, have been:

(segment down-1) location = $1220 \times M + 70{,}425$
(segment down-2) location = $1297 \times M + 154{,}716$

The slope in both equations indicates the average cycle measured in seconds between the adjacent maxima. That cycle is about 21 minutes (1260 seconds).

Physical interpretation of the results

How can we interpret the discovered patterns? Recall that the readings should be constant or fluctuate minimally, as the beaker has not changed its mass.

We do not know what caused the one-second extrema at the highest scale. Perhaps an error in analog-to-digital conversion. But we can interpret many patterns at the lower levels. Consider the linear relation found for maxima locations, *location* $= 86855M - 18261$. It indicates a constant cycle of 86,855 seconds. When compared to 24 hours (86,400 seconds), it leads to an interesting interpretation: the cycle is just slightly longer than 24 hours. The measurements have been made in May, when each day is few minutes longer then the previous one. These facts make us see a close match

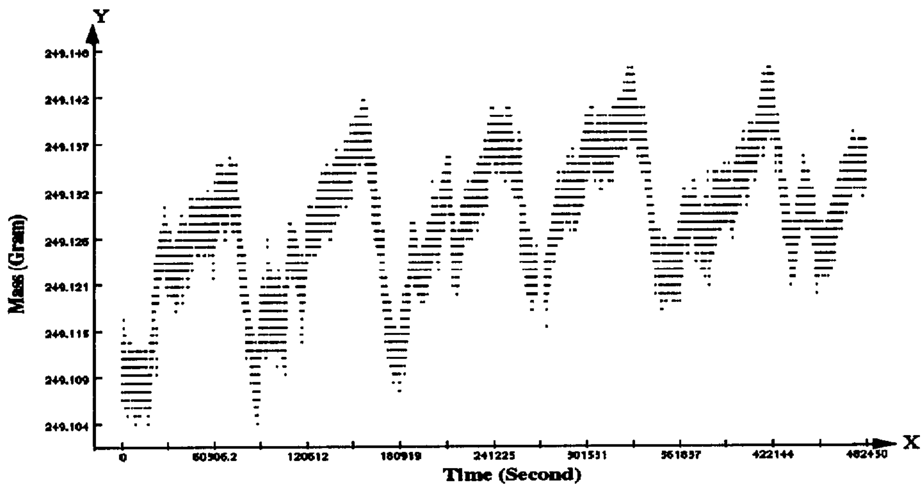

Figure 2: Mass measurements on electronic balance after the outlier extrema have been subtracted.

Scale	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
M_{max}	14337	3114	2308	808	617	267	204	78	46	26	22	16	15	12	12	11	7	7	7	7	6	6

Scale	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43
M_{max}	5	5	5	5	5	5	5	5	4	3	2	2	2	1	1	1	1	1	1	1	1

Table 1: The number of maxima (M_{max}) found for different scale values.

between the cycle of day and night and the maxima and minima in the data. The mass is the highest at the end of day. It goes sharply down through the night, which is much shorter than the day in May. Then the mass goes slowly up during the day, until the next sunset.

Why would the balance reflect the time of the day with such a precision? Among many possible explanations we can consider temperature, which changes in a daily cycle. The actual changes in temperature at the balance did not exceed one centigrade and have been hardly noticeable, but if this hypothesis is right, it should apply to other patterns discovered in the same data. The balance has been located in a room with the windows facing east and the morning sun raising temperature from early morning. What about the short term cycles of about 20 minutes? The air conditioning seems the culprit. It turns on and off about every 15–25 minutes, in shorter intervals during the day, in longer intervals at night. The regularities in maxima locations in different data partitions reveal that pattern. The room has been under the influence of both the air-conditioning and from the outside, which explains both the daily cycle and the short term cycle. We could not confirm these conclusions by direct measurements. Few weeks later our discovery lab has been moved to another building.

Acknowledgment: special thanks to Jieming Zhu for his many contributions to the research reported in this paper.

References

Hoaglin, D.C., Mosteller, F., & Tukey, J.W. 1983. *Understanding Robust and Exploratory Data Analysis*, John Wiley & Sons, New York

Kendall, M. & Ord, K. 1990. *Time Series*, Edward Arnold Publ.

Klösgen, W. 1996. Explora: A Multipattern and Multistrategy Discovery Assistant, in U.Fayyad, G.Piatetsky-Shapiro, P.Smyth & R.Uthurusamy eds. *Advances in Knowledge Discovery and Data Mining*, AAAI Press, 249-271.

Langley, P., Simon, H., Bradshaw, G. & Zytkow, J. 1987. *Scientific Discovery: Computational Explorations of the Creative Processes.* MIT Press: Cambridge, Mass.

Piatetsky-Shapiro, G. and C. Matheus, 1991. Knowledge Discovery Workbench, in G. Piatetsky-Shapiro ed. *Proc. of AAAI-91 Workshop on Knowledge Discovery in Databases*, 11–24.

Tierney, L. 1990. *Lisp-Stat: An Object-Oriented Environment for Statistical Computing and Dynamic Graphics*, Wiley, New York, NY.

Tukey, J.W. 1977. *Exploratory Data Analysis*, Addison-Wesley, Reading, MA.

Witkin, A.P., 1983. Scale-Space Filtering, in *Proc. of Intl. Joint Conf. on Artificial Intelligence*, AAAI Press.

Zembowicz, R., & Żytkow, J.M. 1992. Discovery of Equations: Experimental Evaluation of Convergence, in *Proc. of 10th National Conference on Artificial Intelligence*, AAAI Press, 70–75.

Żytkow, J.M. 1996. Automated Discovery of Empirical Laws, to appear in *Fundamenta Informatica.*