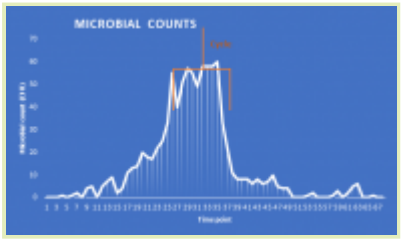
IVT NETWORK
INSTITUTE OF VALIDATION TECHNOLOGY | an informa business
Unlimited Validation Knowledge Awaits...
Published on *IVT Network* (http://www.ivtnetwork.com)

Digital Data #4: Looking for Data Trends and Patterns With Visualization

By **Tim Sandle**

Mar 9, 2022 8:00 am EST



INTRODUCTION

Looking for patterns in data is a key part of many sectors, including those working in pharmaceuticals and healthcare. While software packages are available to automate this task and, for highly complex data, machine learning can be used, reviewing data continues to matter event for the simpler aspects of life in the sectors. This includes assessing metrics from batch record data (such as yield or a critical process parameter); microbial counts; types of deviations by category and so on. For pattern analysis, data visualization provides a quick, easy way to convey concepts universally (1).

This article presents some approaches that can be used in order to obtain insights from data using simple visual tools. While the tools are simple, they can help with conveying an important point with clarity (2).

The article follows on from others in IVT's digital data series:

1. Digital Data #1: Content Creation: https://www.ivtnetwork.com/article/digital-data-1-content-creation
2. Digital Data #2: Digital Search Challenges: https://www.ivtnetwork.com/article/digital-data-2-digital-search-challenges
3. Digital Data #3: Data Governance and the Patient: https://www.ivtnetwork.com/article/digital-data-3-data-governance-and-patient

The continuation of the theme is that provided data can be captured digitally then the possibilities for deconstructing, categorizing, sorting, and presenting are many and these can help value to be drawn from the data analysis step.

WHAT IS A DATA PATTERN?

When seeking to decompose a problem, this involves seeking to find patterns among the data produced. The patterns are similarities or characteristics that a criterion that is shared. More complex pattern recognition is fundamental to computer science. It is also something that can be undertaken with data sets in software like MS Excel or Minitab.

A pattern within a set of data is a series of data that repeats in a recognizable way. This can be identified in the history of the data being evaluated or other data with similar characteristics. The simplest forms of patterns are numbers trending upwards or downwards. These patterns are often clearer when the numerical data is presented graphically or in a table format. Patterns can also be discerned from simple statistics, such as looking for correlations between two sets of numbers.

Two common data patterns are those clustered around time (as with a trend chart) and those centered on causality (as with regression analysis. Time series models assume that the direction the chart takes is only related to its own past patterns; whereas causal models measure the relationship between the other factor(s) and the data being considered.

UNDERSTANDING HOW THE DATA WAS GATHERED

Understanding how the data was gathered and what it relates to is important, especially before looking at the data set for patterns. Often data will fall into two groups:

- Cross-sectional data: These are observations collected at single point in time, for example a series of tests for one in-process sample.
- A time series data: These are collected over successive increments of time. For example, a series of in-process samples examined in relation to one test.

The way the data was gathered enables different patterns to be examined. With time-based data there are typically four general types of patterns: horizontal, trend, seasonal, and cyclical (3). With cross-sectional data the focus is more with extracting information and discerning patterns from individual events.

TIME AND TREND

When sufficient data points exist and the data has been drawn across a selected time period that is meaningful, then a trend can be discerned. The trend is the long-term component that represents the growth or decline in the time series over an extended period of time. Line charts are best for continuous data as it connects many variables that all belong to the same category.

With these types of charts, the data points may vary slightly, but overall, the data moves in one direction. For example, microbial counts increasing for the same sample as measured across a period time are displayed in Figure 1.

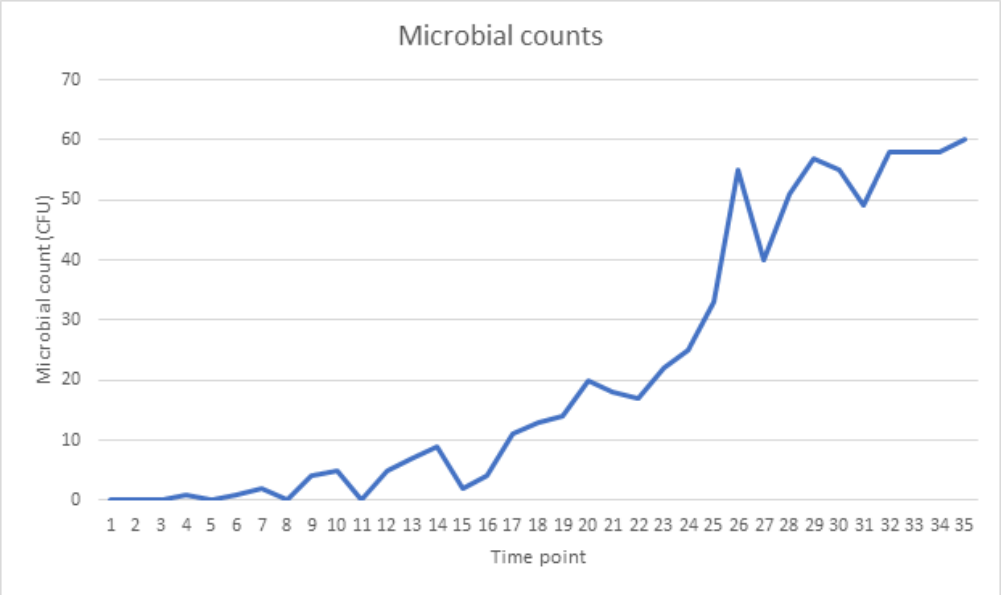


Figure 1: Data examined across time, for microbial counts

Or with pH readings, taken from the same point the process across successive batches, the data is shown to be declining across a series of time points (as per Figure 2).

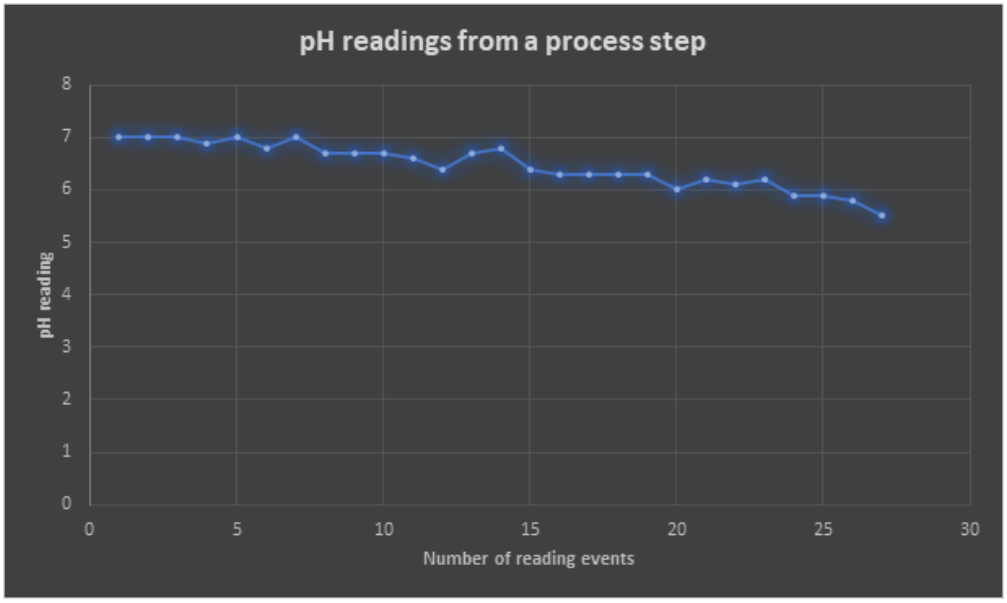


Figure 2: Data examined across time, for pH readings

The time-based data can also be assessed for its cyclical nature. The cyclical component is the wavelike fluctuation around the trend. This can be illustrated by returning to the microbial count data (which subsequently improved as time progressed following some corrective actions).

Here a cycle can be called out, as Figure 3 shows.

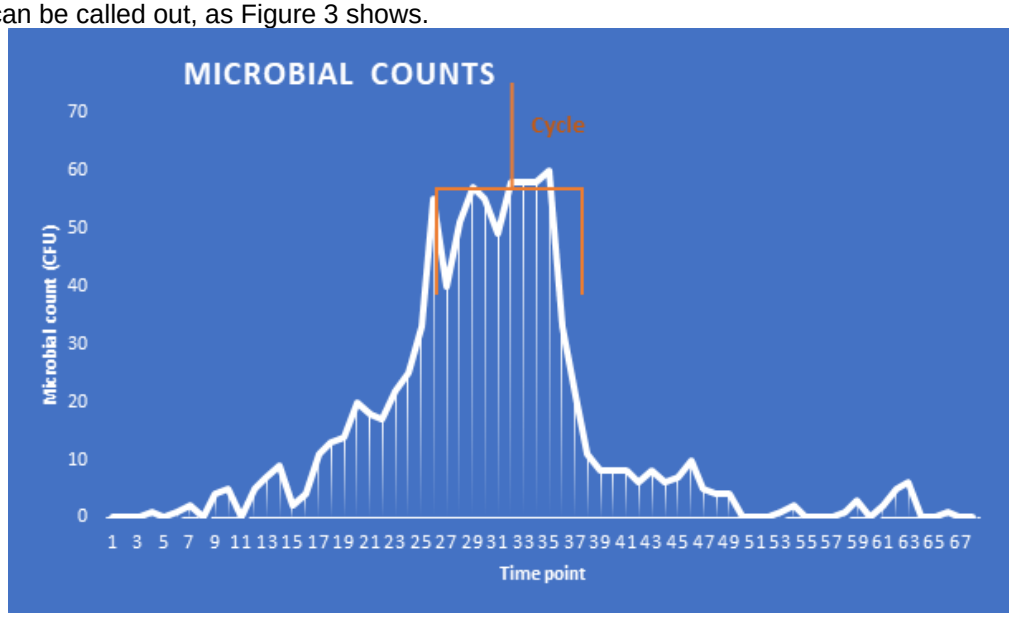


Figure 3: Zoning in on a data cycle (in relation to microbial counts)

Cycles can relate to other events, such as periods within the overall time period of the graph, such as a month or quarter within a year, or in relation to a controlled change. Cycles can also appear as long-wave patterns, and these sometimes repeat. Take, for example, microbial counts from a water system, as per Figure 4.

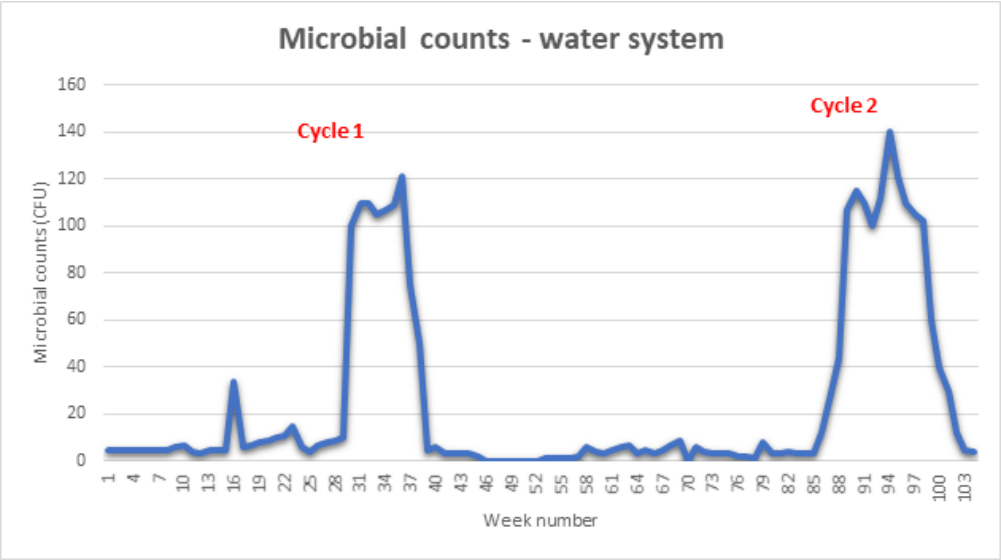


Figure 4: Illustration of a repeating cycle in terms of seasonality, for microbial counts

Here there are two evident cycles across two years coinciding at approximately the same time of the year. This might suggest a period of time (the summer) when microbial counts elevate. In the hypothetical situation, the increases in counts are linked to a period of production shutdown, for maintenance reasons. Often such cyclical patterns are regular, although they can vary in length.

The cyclical pattern may lead into the seasonal component, which is a pattern that repeats itself year after year.

When data collected over time fluctuate around a constant level or mean, a*horizontal pattern* exists. This type of series is to be stationary in its mean. Monthly yields for an active pharmaceutical ingredient that do not increase or decrease consistently over an extended period would be considered to have a horizontal pattern.

EXPLAINING THE PATTERN

When collected data is examined, it is useful to be able to add a descriptor to the data:

- Are the data random? (Such as where successive values of a time series are not related to each other)
- Do the data have a trend (are they nonstationary)?
- Are the data stationary or horizontal?
- Are the data seasonal?

With the above, a series that varies about a fixed level (no growth or decline) over time can be described as stationary. Hence, a stationary time series is one whose basic statistical properties, such as the mean and variance remain constant over time. Whereas a series that contains a trend can be said to be non-stationary.

DISTRIBUTION

Another way of looking at data is in terms of its distribution. For this, graphic displays are useful for seeing patterns in data. This form of visual analysis can be used throughout a study to make informed decisions or changes about design and study variables while maintaining experimental control and producing improved outcomes. The analysis can also be useful for examining data for normalcy (or otherwise) prior to selecting the appropriate statistical tool to use.

Patterns in data distribution are commonly described in terms of center, spread, shape, and unusual features (4). When plotted, where central data is observed, the center of a distribution is located at the median of the distribution (refer to Figure 5). This is the point in a graphic display where about half of the observations are on either side. In the chart below, the height of each column indicates the frequency of observations. This represents a form of normal distribution.

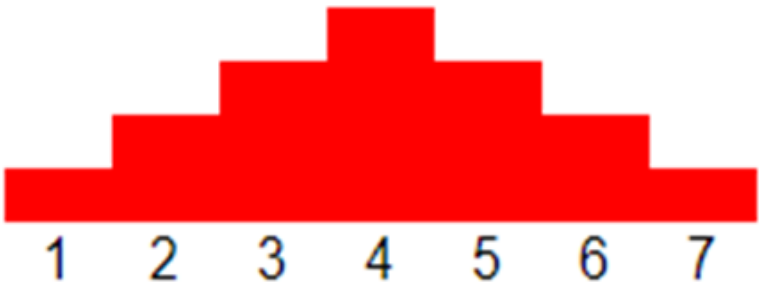


Figure 5: Data showing central distribution

This is also coincidental with symmetry. When it is graphed, a symmetric distribution can be divided at the center so that each half is a mirror image of the other. Number of peaks. Distributions can have few or many peaks. Distributions with one clear peak are called unimodal, and distributions with two clear peaks are called bimodal. When a symmetric distribution has a single peak at the center, it is referred to as bell-shaped. This is not the

same as uniform distribution, which is when the observations in a set of data are equally spread across the range of the distribution (as per Figure 6). A uniform distribution has no clear peaks.

Figure 6: Data showing uniform distribution

**Images based on:https://stattrek.com/statistics/charts/data-patterns.aspx*

The spread of a distribution refers to the variability of the data. If the observations cover a wide range, the spread is larger (as with Figure 8). If the observations are clustered around a single value, the spread is smaller (as with Figure 7).

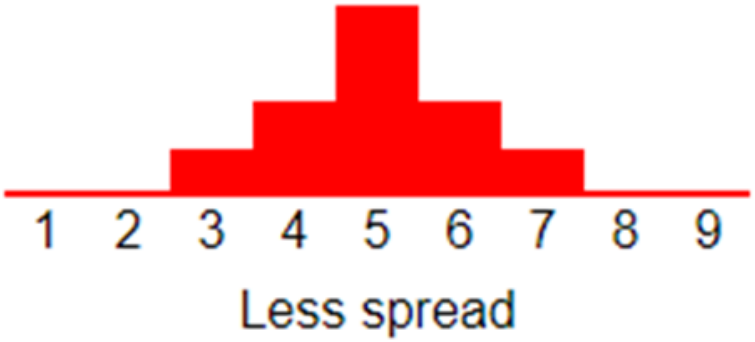


Figure 7: Data where the spread is limited

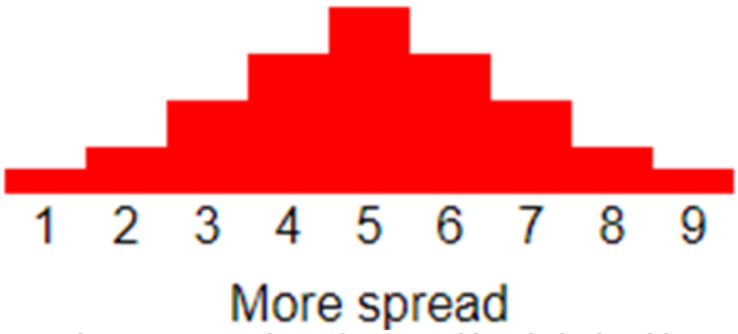


Figure 8: Data where the spread is relatively wide

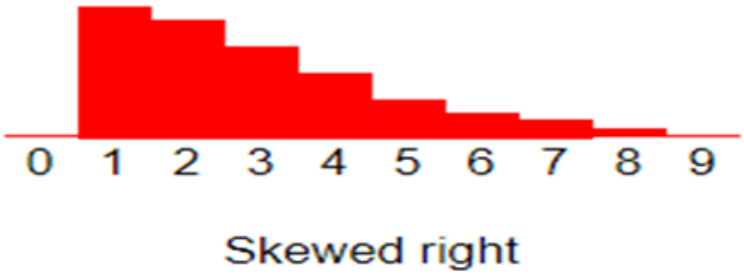


Figure 9: Data displaying right-hand skewness as if often the case with microbial data

Skewness. When they are displayed graphically, some distributions have many more observations on one side of the graph than the other. Distributions with fewer observations on the right (toward higher values) are said to be skewed right; and distributions with fewer observations on the left (toward lower values) are said to be skewed left. Microbiological data, for example, tends to exhibit right skewness (as per Figure 9).

Outliers. Sometimes, distributions are characterized by extreme values that differ greatly from the other observations. These extreme values are called outliers. The figure below illustrates a distribution with an outlier. As a "rule of thumb", an extreme value is often considered to be an outlier if it is at least 1.5 interquartile ranges below the first quartile (Q1), or at least 1.5 interquartile ranges above the third quartile (Q3). An example of such an outlier is shown in Figure 10.

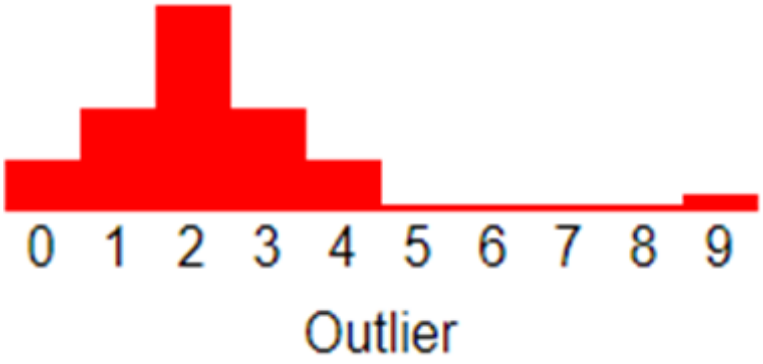


Figure 10: Distribution chart showing the presence of an outlier value. In such cases a case could be made to exclude the outlier from subsequent analysis.

OTHER CHARTS

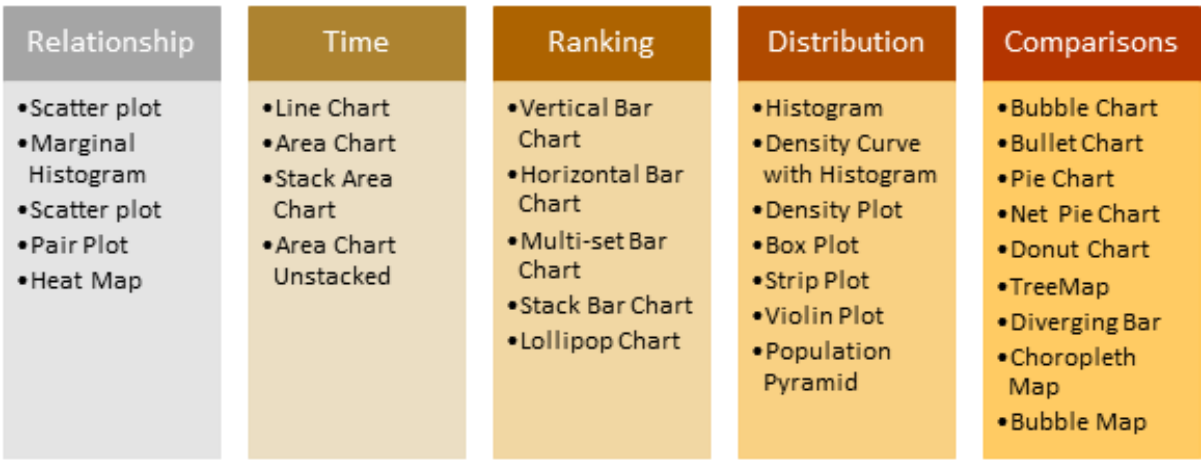


Figure 11: Image showing the rage of different charts available for conducting data pattern analysis for visual purposes.

When assessing a relationship between data sets, the objective is to understand how two or more data sets combine and interact with each other. This relationship is called correlation, and it can be positive or negative, meaning that the variables considered might be supportive or working against each other. One effective way to do this is using a scatterplot.

For data ranking, the simplest method is with the bar chart, composed of a series of bars illustrating a variable's development. There are four types of bar charts: horizontal bar chart, vertical bar chart, group bar chart, and stacked bar chart.

TABLES

When data is tabulated, often the data is categorized or placed into range. To aid this, sorting and filtering are common tools that allow for data to be organized. The sorting of data is concerned with putting it into an order; whereas, with filtering data, this allows unimportant data to be hidden and for the user to focus only on the data they are interested in.

HOW DATA SETS GO WRONG

Looking for data patterns is meaningless if the data itself is unsuitable. Hence, an assessment should be made as to the source and representativeness of the data. This will also include ensuring the data set is sufficiently large. To ensure that data is suitable, the data should be assessed to see:

- Is the data should be reliable and accurate?
- Is the data relevant?
- Is the data representative of the circumstances for which they are being used?
- Is the data consistent?
- Was all of the data collected under the same definition?
- Does any element of the data require adjusting to retain consistency in relation to historical patterns?
- Does the data for a suitable time period?
- Has sufficient data bene included?

MACHINE LEARNING

More sophisticated data pattern recognition exists in machine learning. The machine learning algorithm learns from data and once optimized, it will automatically recognize patterns even if partially visible. While the process will recognize familiar pattern, the recognition comes from different shapes and angles, and this is where the

sophistication provide by machine learning can prove to be very useful (5).

SUMMARY

This article has looked at some simple data presentation tools, including ways to capture and sort data, and to look at data across time, as a correlation, or in terms of its distribution. Of course, there are may other approaches and more complex inquires can be undertaken. The purpose here was not to be comprehensive; instead, it was to offer a few examples for those setting out on the data review journey. In doing so, the focus has been on visual representation of data rather than the examination of data through statistical analysis. Often, looking at visual pictorial can reveal a great deal about the shape or feel of the data. This may be enough for the inquiry at hand, or it may pave the way for statistical assessment.

REFERENCES

1. F. Afrati, A. Gionis, and H. Mannila. Approximating a collection of frequent sets. In*Proc. ACM SIGKDD*, 2004
2. Ajani, K., Lee, E., Xiong, C., Knaflic, C. N., Kemper, W., Franconeri, S. (2021). Declutter and focus: Empirically evaluating design guidelines for effective data communication. *IEEE Transactions on Visualization and Computer Graphics*. https://doi.org/10.1109/TVCG.2021.3068337
3. Ancker, J. S., Senathirajah, Y., Kukafka, R., Starren, J. B. (2006). Design features of graphs in health risk communication: A systematic review.*Journal of the American Medical Informatics Association*, 13(6), 608–618
4. Chance, B., delMas, R., Garfield, J. (2004). Reasoning about sampling distributions. In Ben-Zvi, D., Garfield, J. (Eds.),*The challenge of developing statistical literacy, reasoning, and thinking* (pp. 295–323). Springer.
5. C. M. Velu and K. R. Kashwan, "Visual data mining techniques for classification of diabetic patients," 2013 *3rd IEEE International Advance Computing Conference* (IACC), 2013, pp. 1070-1075, doi: 10.1109/IAdCC.2013.6514375.
