TECHNION – Israel Institute of Technology
Faculty of Industrial Engineering & Management

Center for Service Enterprise Engineering (SEE)

SEELab https://seelab.net.technion.ac.il/

SEE Software and Data

SEEStat 3.0+ Tutorial

Exploratory Data Analysis (EDA) of Service Systems (Call Centers, Hospitals)

Created: June 27, 2010
Last Updated: November 8, 2020

SEEStat 3.0+ Tutorial

Introduction

SEEStat is a software platform for Exploratory Data Analysis (EDA) in real-time. It enables users to easily conduct statistical and performance analyses of massive datasets; in particular, analyzing datasets that represent operational histories of large service systems, such as those available through the SEELab server (e.g. call centers, hospitals and in particular emergency departments, internet websites). SEEStat can also automatically create sophisticated reports in Microsoft Excel, which summarize EDAs and hence support research and teaching.

Both SEEStat and the SEELab Server were developed at the Faculty of Industrial Engineering and Management, Technion, Israel Institute of Technology. More information on the SEELab can be found at its homepage.

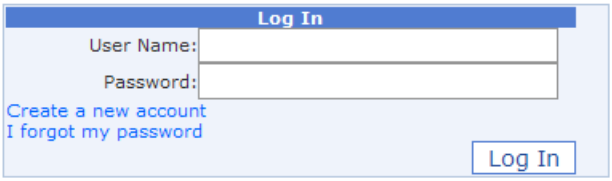
Connecting to SEEStat on the Technion SEELab Server

NOTE: The following assumes that you already have registered successfully to SEELab. In particular, this entails that you now have working User-Name and Password.

The standard way to connect to the SEE server is via the **Microsoft Internet Explorer web browser <u>only</u>,** within **all versions of Windows (starting from Windows 7)**.

1. From Internet Explorer visit this address:
   https://see-center.iem.technion.ac.il/terminal-see/
   (You may wish to bookmark this URL for future use.) You will see the following:

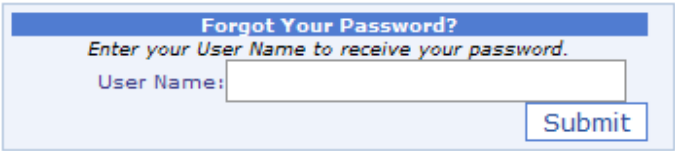
2. Select **"Log In"**, type your User Name and Password, and then click the **"Log In"** button.



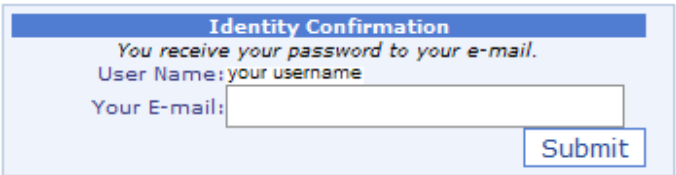
If the User Name and Password are valid, you will have access to the SEE terminal and proceed to **Step 4**. If you forgot your password – proceed to **Step 3**.

3. Click link **"I forgot my password"** in window **"Log In"**.

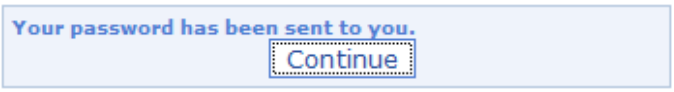
   3.1 Type your User Name and click button **"Submit"**.



   3.2 Type your e-mail and click button **"Submit"**.



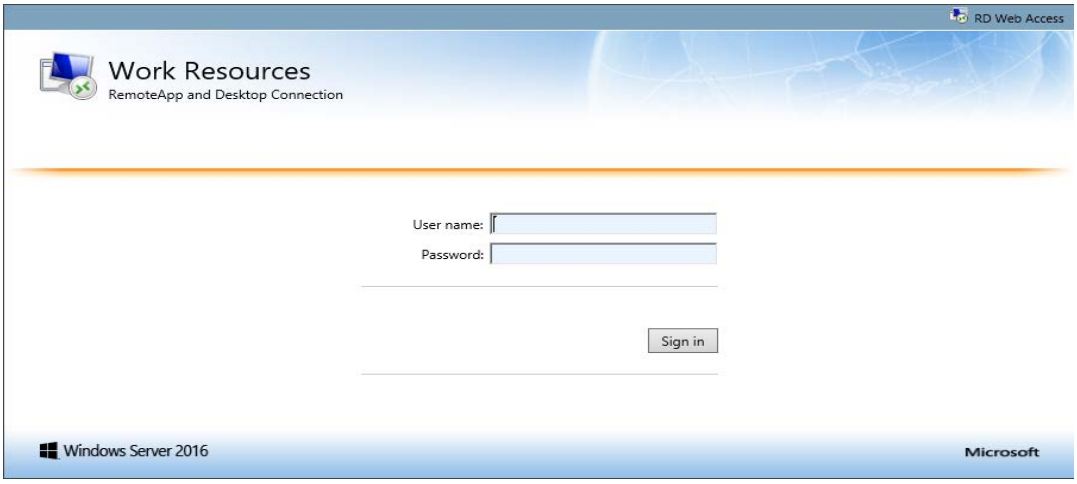
3.3 Click button **“Continue”**.



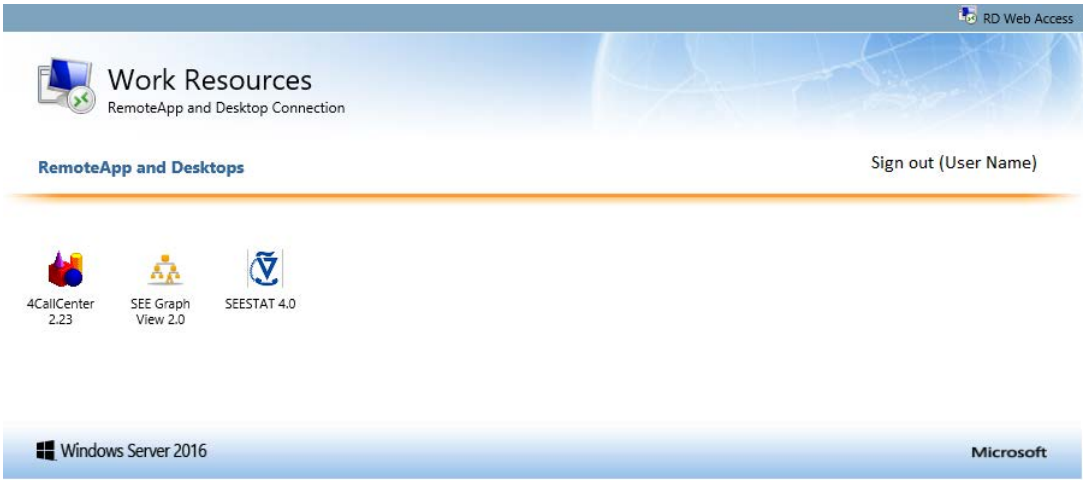
Over few minutes, you receive e-mail containing following information: your User Name and your Password.

4. Click “**To Terminal**"

After redirecting to below window, type your User Name and Password from **Step 2**, and then click button **“Sign In”**.

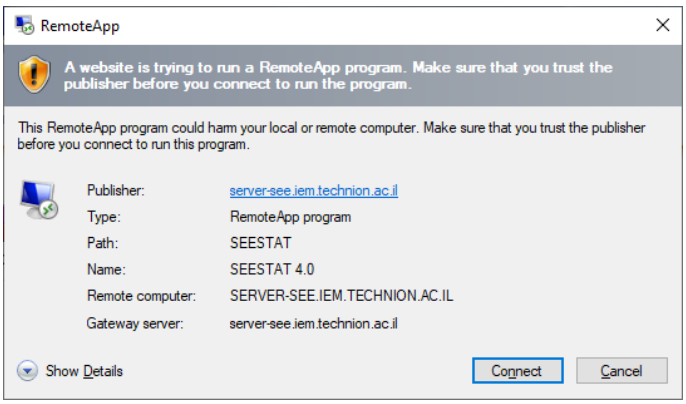


5. After verification, you will have access to window with programs.



6. Run program.

6.1 Click the SEEStat icon to run the program and after click **“Connect”**.

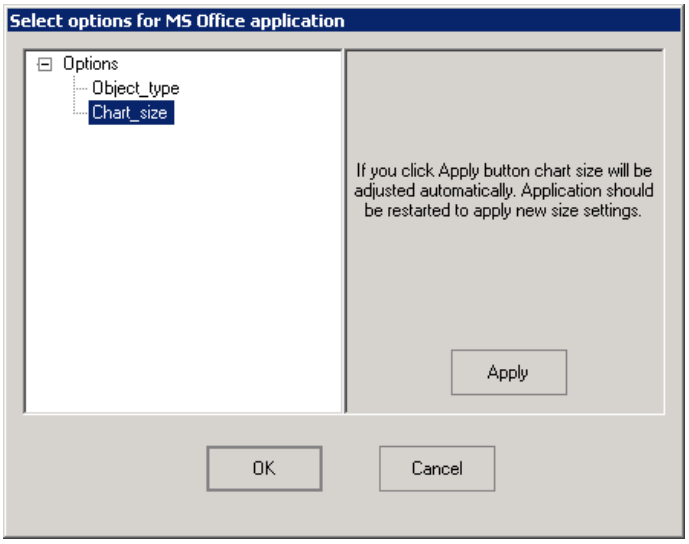


If you not see this window, please read step 7

SEEStat interacts with Excel to display data. You might then discover that the Excel chart size does not fit your screen size. Read in 6.2 how to overcome this problem.

6.2 Excel chart size set to fit screen size.

If the Excel chart does not display in full screen click **Output→Options**. Select option **Chart_Size**. Click **Apply** and **OK**.
Commentary: This option is used in order to reset the chart size into a full screen chart. Typically, this problem does not arise: users will encounter chart size in Excel that does fit their physical screen size. But sometimes, for example, due to a changed resolution, the chart size will not fit screen size. Then, one should use the above option, which looks as follows:



Recall: For the above change to apply, one must exit SEESTAT and re-enter it again.

7. Problem with *Remote Desktop ActiveX* control.

7.1 Add https://see-center.iem.technion.ac.il and https://server-see.iem.technion.ac.il to the Trusted Sites of Internet Explorer.

This is performed as follows: From the Internet Explorer menu, click **Tools → Internet Options**, then visit the **Security Tab**. Select the **Trusted Sites Zone**. Click on **Sites** and **add** the above *URL* to the list of websites.

7.2 Make sure that Internet Explorer has the SEEStat **ActiveX control** enabled.

This is performed as follows: From the Internet Explorer menu, click **Tools → Internet Options**, then visit the **Programs Tab**. Select the **Manage add-ons**. Lookup **MsRdpClientShell** or **Microsoft RDP Client Control** and enable it.

7.3 If **Step 7.1** and **Step 7.2** do not help, send e-mail to adminsee@technion.ac.il.

Your e-mail must contain the follow information:

a. Your operation system (Windows…).
b. Your Web Browser and version.
c. Your problem.

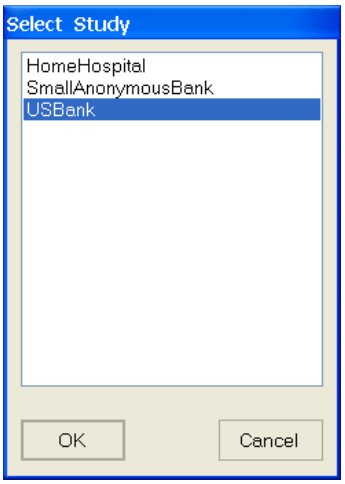
8. Disconnecting from the SEELab Server.

To end your Remote Desktop session: click **Sign Out**.

SEEStat Tutorial

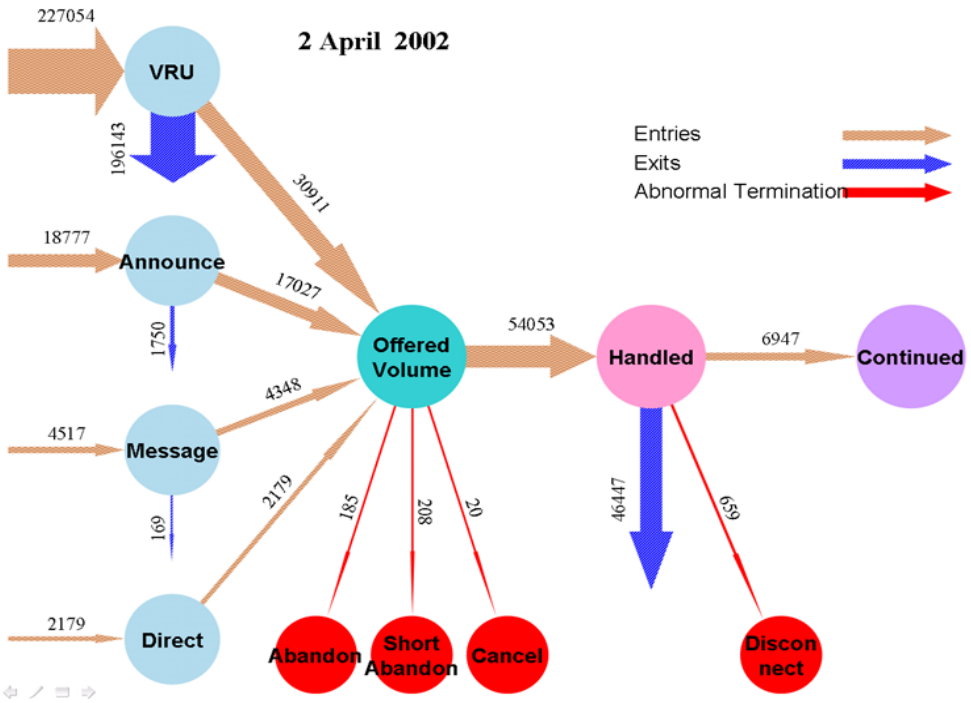
USBank Data

Background: The source of this example database is a large call center of a U.S. Bank. This call center has sites in 4 states, which are integrated to form a single virtual call center: Calls are queued up, when appropriate, in a central queue; they are then served by agents across sites, by fitting service types to agent skills using SBR (Skills-Based Routing) algorithms.
The virtual call center has about 900–1200 agent positions on weekdays, and 200–500 agent positions on weekends. Agents process up to 300,000 calls per day (about 20% reach the agent-queue, and the rest complete their service process within the VRU = Voice Response Unit).



Customer flow chart of the USBank call center

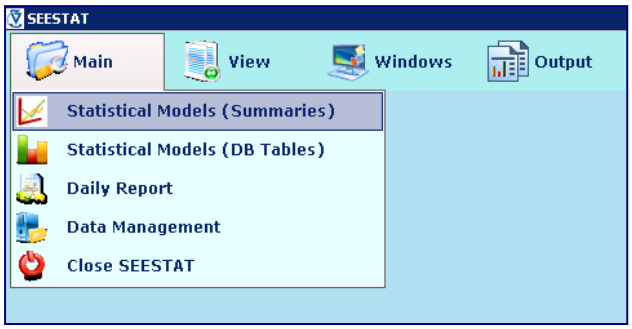
The following flow-chart describes the process-flow of calls in a typical day (Tuesday, April 2, 2002). There are 4 entry points to the system: through the VRU (Voice Response Unit), Announcement, Message, and Direct group (callers that directly connect to an agent). The most commonly used is the VRU. Most of the calls end service in the VRU (196143 calls - about 80% of all calls); while around 20% of the callers entering the system seek service by an agent ('Offered Volume'). Less than 1% of the Offered Volume calls will not reach an agent service – those customers abandon the queue while waiting (a few are disconnected due to technical problems – 'Cancel'). When an appropriate agent becomes available the customer is getting served ('Handled'), after which the customer either exits the system or is moved to a secondary service ('Continued').



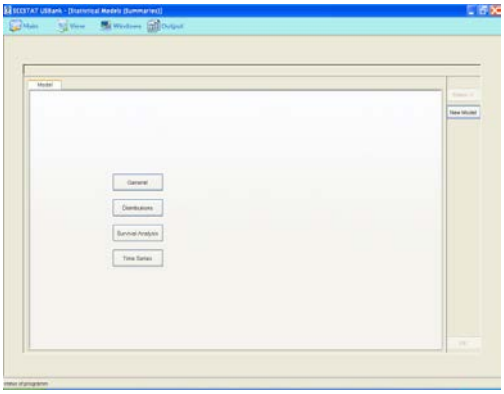
NOTE: Creating such flow-charts is very easy in SEEStat. (Indeed, when you are done going over the tutorial, you will reach Appendix A – it starts with a short guide on few button-pushes that create the above chart as a PowerPoint file.)
Also note that this could be your first step of going beyond flow-charts, and towards creating data-animations: SEEStat has a relative, SEEGraph, which enables semi-automatic creation of data-animations that are far more data-intensive, complex and insightful. (The above-mentioned Appendix A constitutes a brief introduction to SEEnimations.)

Part 1

After connecting to the server, click the SEEStat 3.0 icon to open the program.
On the top of the screen you see the main menu. Click **"Main"**. We shall work with **"Statistical Models (Summaries)".** Click it.



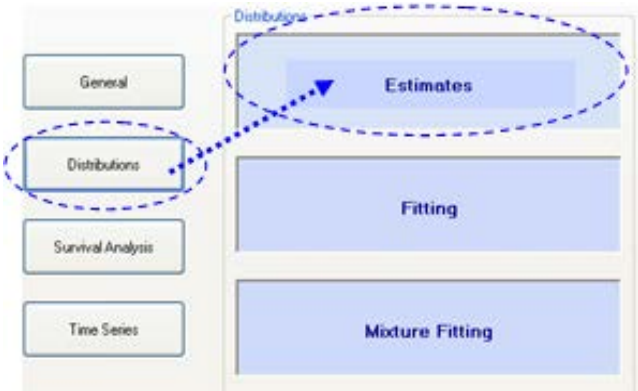
A list-box with SEEStat *studies* appears (three databases in our case). Select **USBank** (the database we shall be working with), click **"OK"** and wait a few seconds.
Now you see the **"Model"** panel.



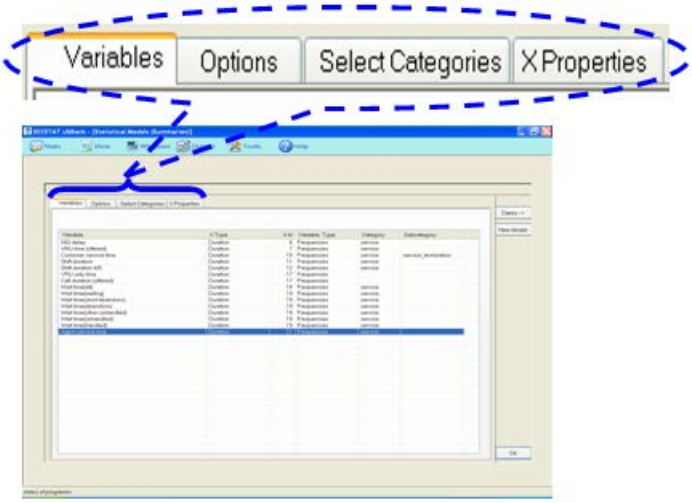
<u>Example 1.1:</u> Distributions

We shall now create a histogram of the ***service time*** *(duration) distribution, at 1-second resolution.*

Click the **"Distributions"** button. Three available distribution models appear.
Select **"Estimates"**.



You see the tab control that has 4 tabs:
“Variables”, “Options”, “Select Categories” and “X Properties”.

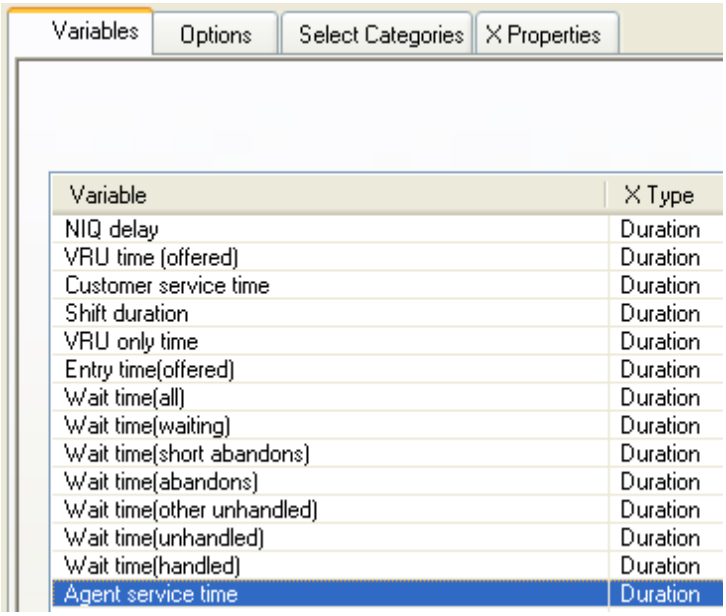


The first one **"Variables"** is active. This tab is *mandatory*, which means you must select variable(s) before moving forward. The three other tabs are *optional*, which means that they already have default values.

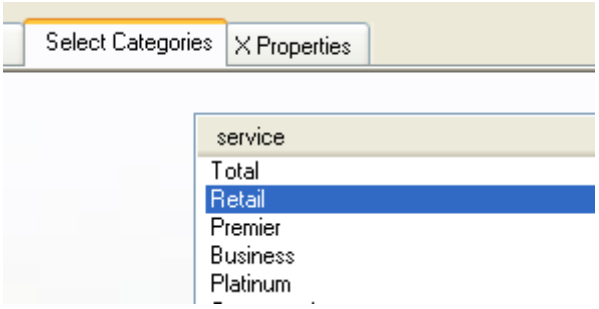
NOTE: You can select (click) several variables simultaneously by holding the Ctrl key and clicking on the variables one by one.

Select "**Agent service time**"(the last entry in the list).

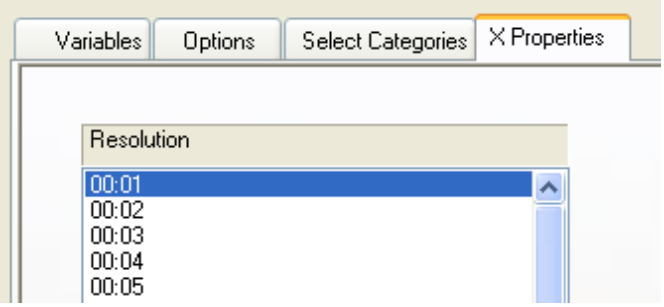
***Remark**: SEEStat provides online definitions for (most of) the variables that it uses: click **View**-> **Summaries (Variables)** and select variable (a demo appears in Appendix F).*



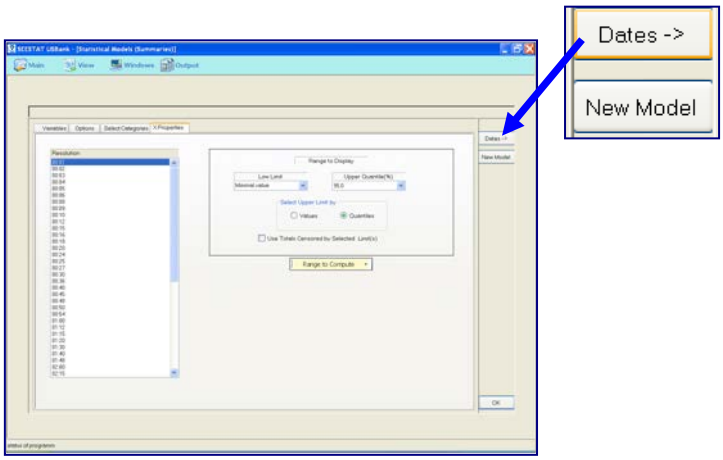
Now move to the **"Select Categories"** tab. You see a list box with all the service types that are offered by USBank. Select **"Retail"**, which is the Bank's main service.



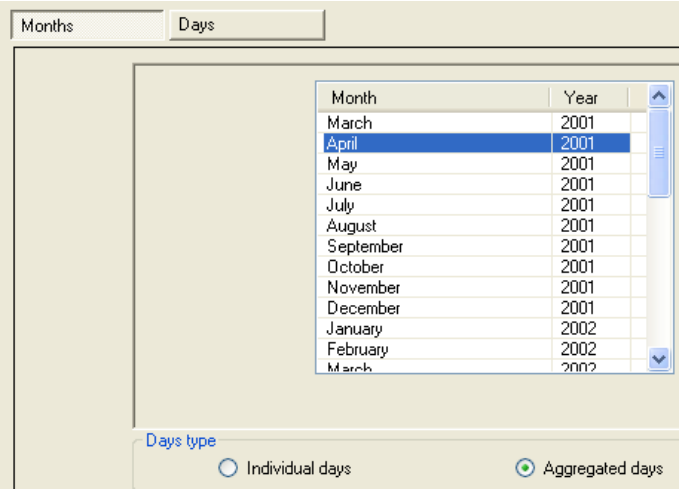
Open the **"X Properties"** tab. It is used to set properties of charts and tables. On the left side you can see the **"Resolution"** list box. The default resolution (bin-size of the histogram) of 5 seconds is marked. Select the minimal resolution **00:01** = 1 second, in order not to miss any details of the histogram.



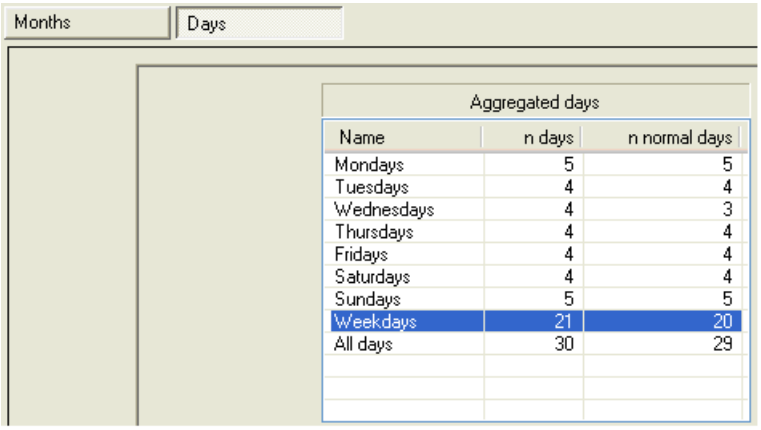
Now you must select the dates we focus on. Click the **"Dates ->"** button on the right side.



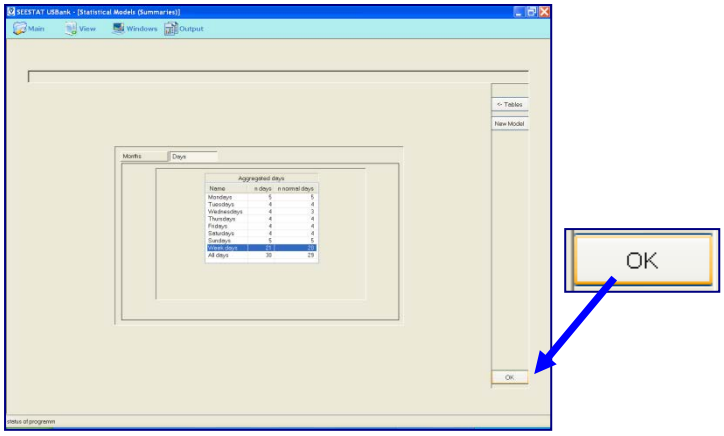
You see the list of months for which the requested data is available.
Select **"April 2001"**.
Below the list of months, you see two options for date-selection (Date type): "Aggregated days" and "Individual days". **"Aggregated days"** is the chosen-default, which we now follow.



Click **"Days"** to make the selection of days, and select **"Weekdays"** – an aggregation of all *5 working days of the week. (Holidays and some special days, such as when there is a system failure, are excluded.)*



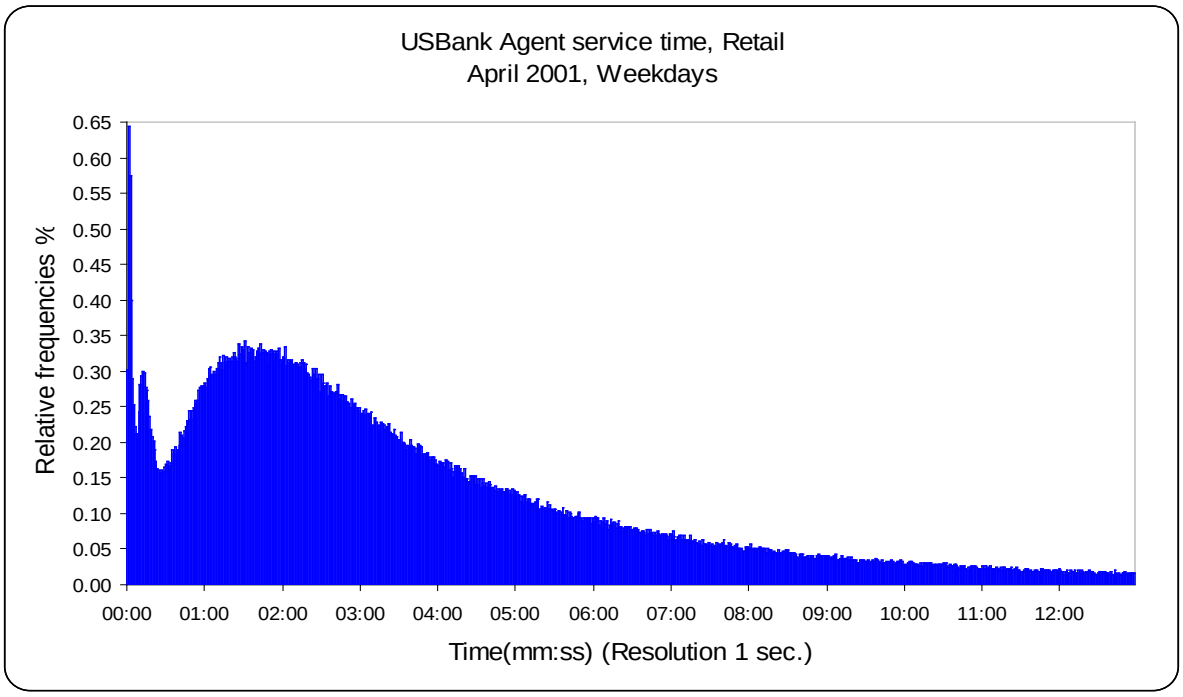
All selections have now been completed: click **"OK"** at the bottom right.



Wait a few seconds *– SEEStat is processing your request: you now see the chart/histogram, produced as "Chart 1" within an Excel spreadsheet.*

NOTE: All the examples in this tutorial, from now on, will be accumulated in this Excel file. It is possible to modify this file (e.g. playing with its graphs), or even close it (e.g. when taking a pause from working on this tutorial). However, one must be then careful to start afresh the relevant example, since some click-history is accumulated. (In particular, if you quit now, you will have to start from the beginning of Example 1.1.)

Looking at the chart, you see some irregularities on the left (near the origin). We shall look at these more carefully later.



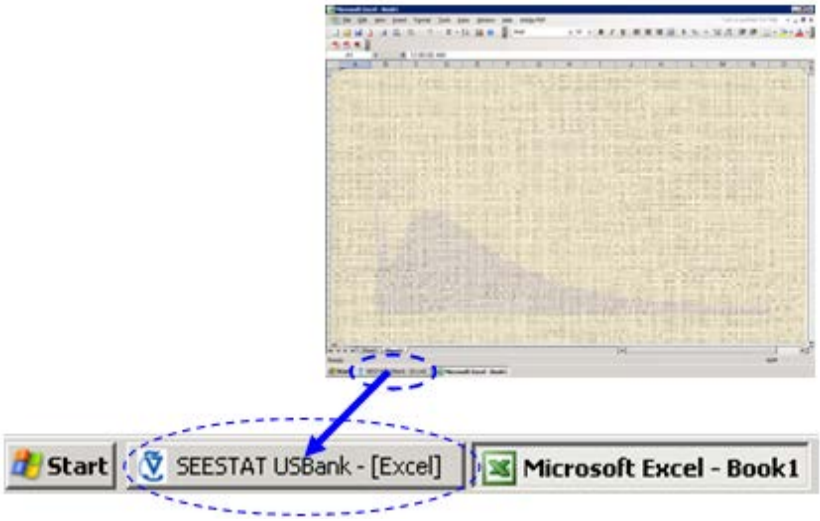
In fact, two sheets have been created: The first is the chart in "Chart1"; the second is "Table1", which includes Table(s) that are associated with the chart, with the default one being the ***"Statistics"*** *table. Click* ***"Table1"*** *to see the contents of this table (N=619,096 is the number of observations); skim through the summary data and then return to* ***"Chart1"****.*

Statistics	
	Agent service time
N	619096
N(average per day)	30954.8
Mean	4 min 19 sec
Standard Deviation	4 min 35 sec
Variance	21 min^2
Median	2 min 57 sec
Minimum	0
Maximum	59 min 53 sec
Skewness	3.062

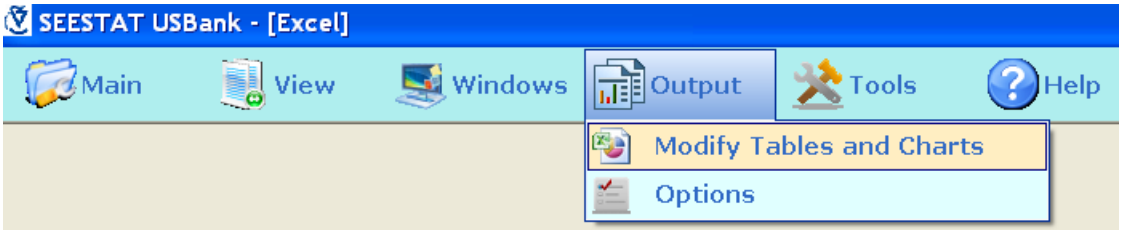
Kurtosis	15.38
Standard Error Mean	0 sec
Interquartile Range	3 min 53 sec
Mean Absolute Deviation	3 min 2 sec
Median Absolute Deviation(MAD)	1 min 42 sec
Coefficient of Variation (CV) (%)	106.17
L-moment 2 (half of Gini's Mean Difference)	2 min 6 sec
L-Skewness	0.383
L-Kurtosis	0.245
Coefficient of L-variation (L-CV) (%) (Gini's Coefficient)	48.57

You can easily make modifications to charts and tables, as long as they do not require the loading of new data from the database. (With new data, you must start a New Model, as will happen many times in the sequel.)
You will now go through an example of such a modification.

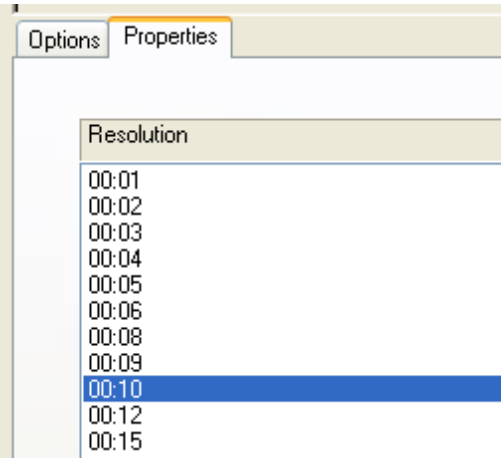
First, return to the SEEStat main menu by clicking the SEEStat button (with SEELab's LOGO), in the task bar on the lower-left side of the screen – <u>you will repeat this action each time you wish to transfer from Excel to SEEStat</u>.



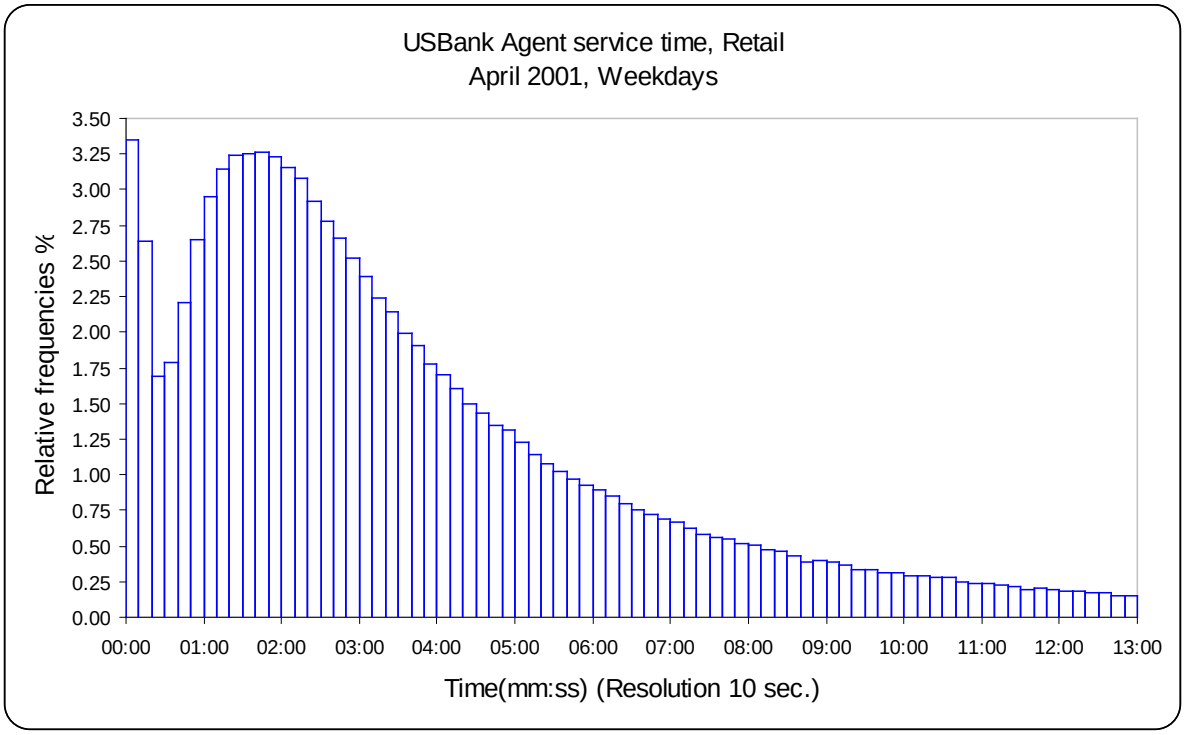
Click **"Output"** on the right side of the top main menu; after this click **"Modify Tables and Charts"**



Two tabs are available: "Options" and "Properties". Open **"Properties"** and change the resolution to **00:10** = 10 seconds.



Click **"OK"**.

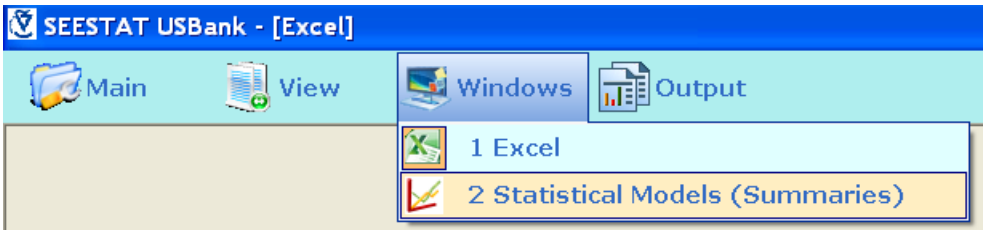


The chart is becoming smoother, but at the cost of losing some details on the left, near the origin.

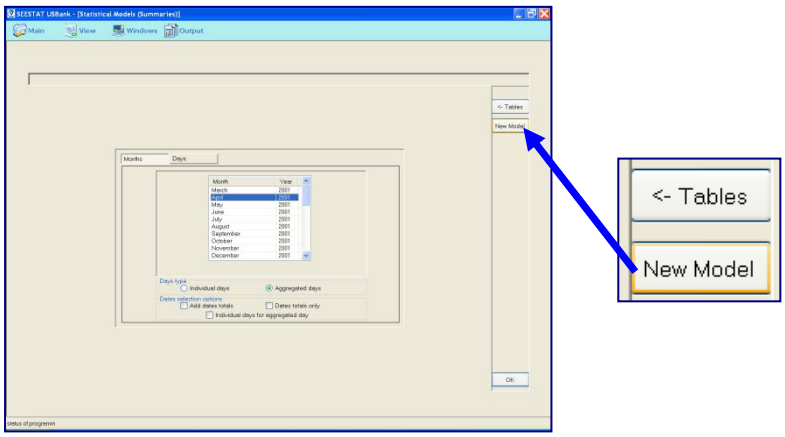
<u>Example 1.2:</u> Intraday time-series

We now create a chart of arrival-counts to the call center(s) of USBank, during several days in a September.

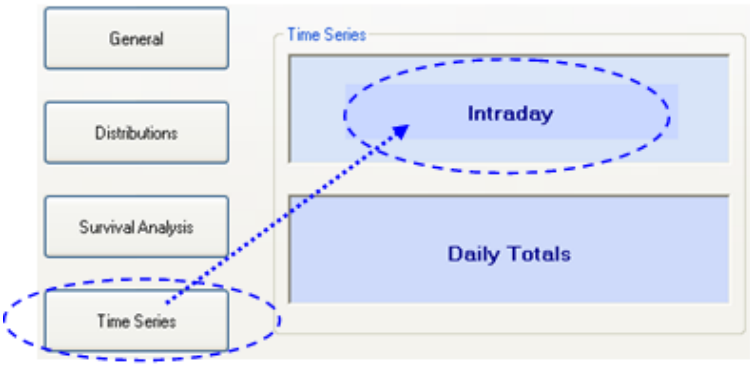
First you must return to the **"Statistical Models (Summaries)"** window. Click the **SEEStat** button on the task bar (left-bottom), next click **"Windows"** on the main menu (at the top) and select **"Statistical Models (Summaries)"**



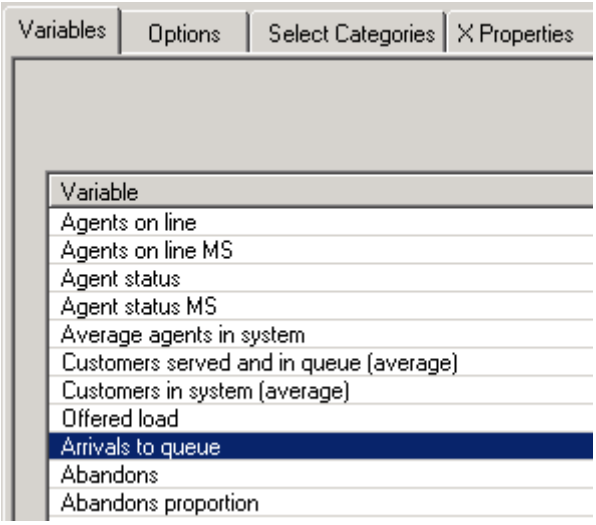
We are now changing models. To this end, select the **"New Model"** button (right side).



Select now **"Time Series"** and then select **"Intraday".**

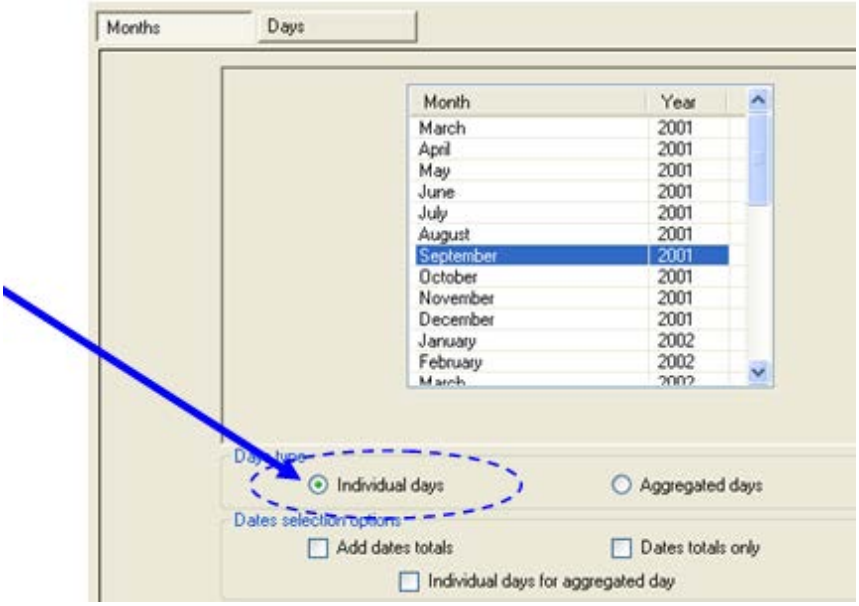


As in Example 1.1, four tabs appear. In the **"Variable"** tab, select **"Arrivals to queue"**. In the **"Select Categories"** tab, select **"Total"**.



Now select dates: Click **"Dates ->"**; Select **September 2001** from the **"Months"** list; Mark **"Individual days"**, and click the **"Days"** button.

.



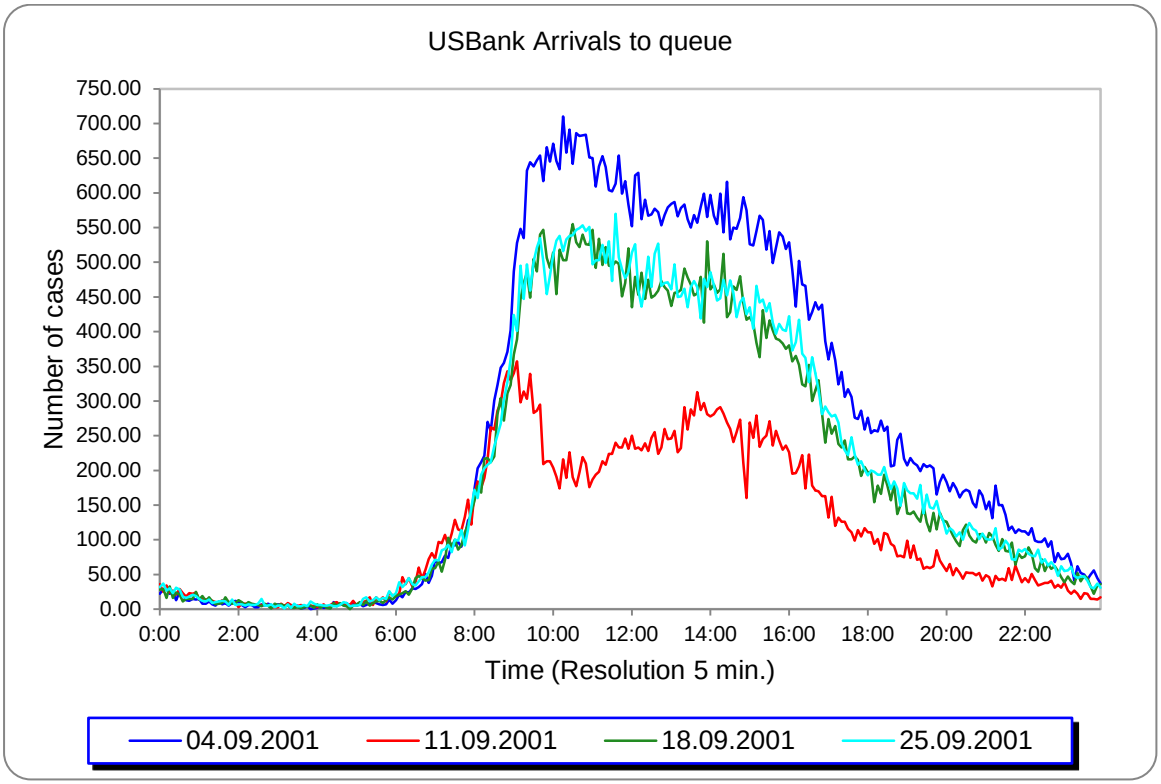
The list of days contains the date, the day of the week and comments if any. For example, Monday, September 3rd, was Labor Day.
It is expected that the Tuesday following a holiday will be a busy day. We thus compare all Tuesdays of the month: September 4, September 11, September 18 and September 25.

Hold down the **"Ctrl"** key, and in parallel click, one by one, the **four Tuesdays** of September 2001.

Days

Day	Month	Year	Week Day	Comments
1	September	2001	Saturday	
2	September	2001	Sunday	
3	September	2001	Monday	Labor Day
4	September	2001	Tuesday	
5	September	2001	Wednesday	
6	September	2001	Thursday	
7	September	2001	Friday	
8	September	2001	Saturday	
9	September	2001	Sunday	ShutDown from 6:3(
10	September	2001	Monday	
11	September	2001	Tuesday	
12	September	2001	Wednesday	

Then click **"OK"** (bottom right).

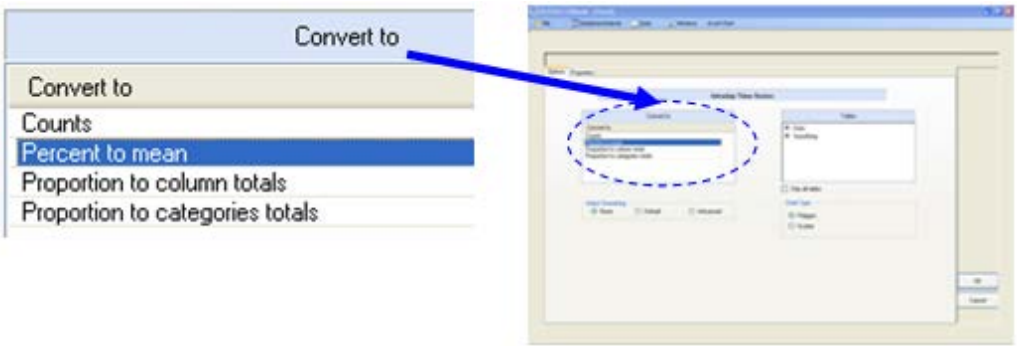


Note: The graphs appear in "Chart2" of Excel. As before,"Table2" contains the corresponding numerical data.

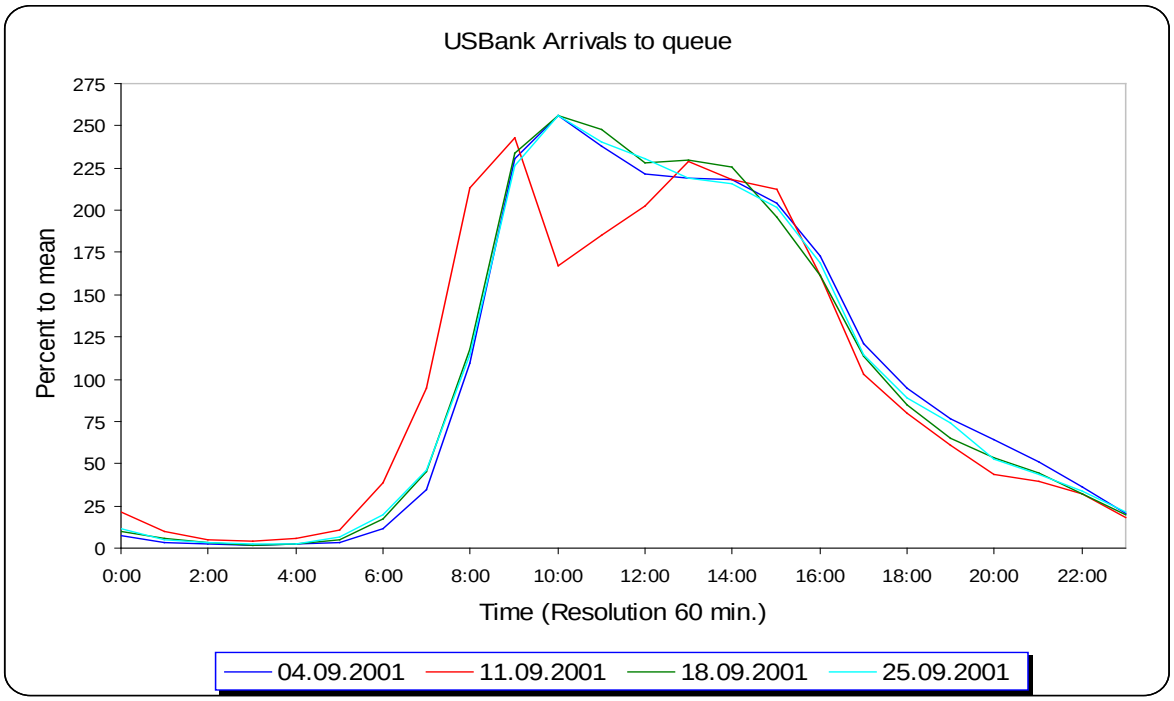
You see a sharp drop in the number of calls around 09:00 a.m. on ***September 11, 2001*** *– this is of course not surprising, recalling the tragic September-11th, and given that one of the call centers of US Bank was in NYC and the others located on the East Coast of the USA.*
You also observe that the Tuesday after Labor Day (4.9) is indeed a heavily-loaded

Tuesday, as anticipated.
The chart is noisy, due to its 5-minute resolution. We shall momentarily increase the resolution to 1 hour (60 minutes). We also note the following:
On the two Tuesdays after September 11, the number of calls is low, relative to the Tuesday after Labor Day. A natural question now arises: Is there a "shape of a Tuesday"? To seek a common pattern for (the shape of) a Tuesday, if there is any, we change the graphs from absolute counts to "percent to mean" (mean = average number of calls per resolution period, which is 5 minutes here).

Go back to the main menu via the **SEEStat** tab (bottom-left). In the main menu select **"Output"** then **"Modify Tables and Charts"**. In the **"Options"** tab, under the **"Convert to"** table on the left, select **"Percent to mean"**, and in the **"Properties"** tab set resolution to **60:00** = 1 hour,



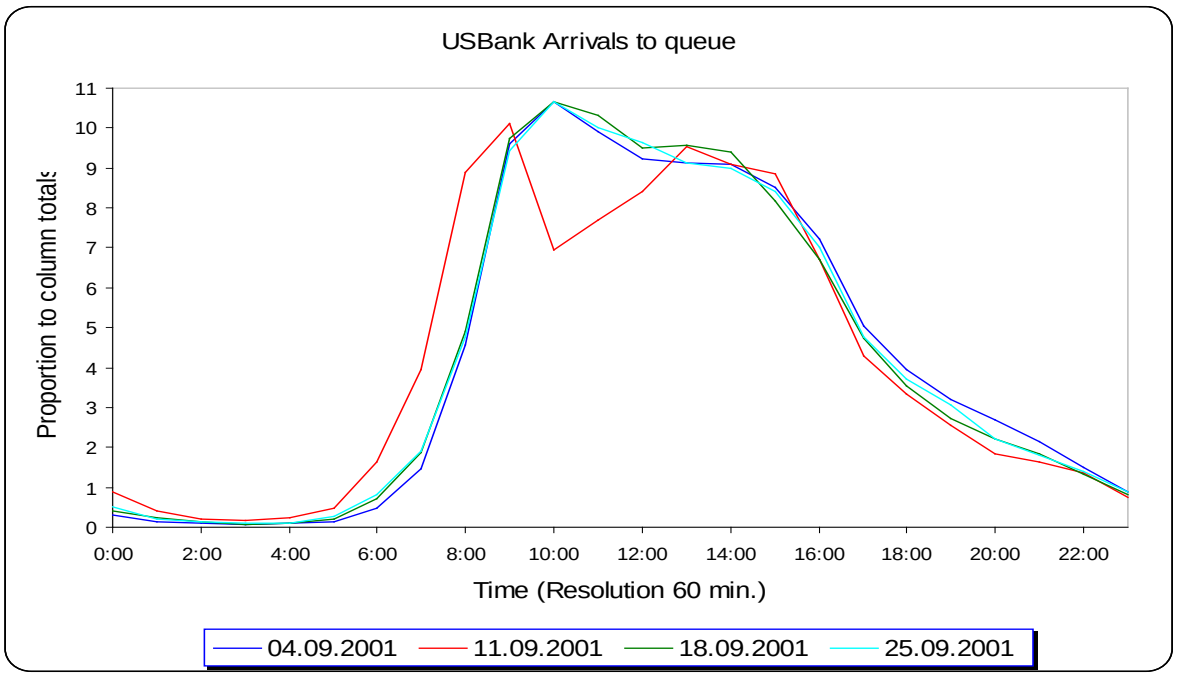
Click **"OK"**.



The "Shape of a Tuesday" is clearly manifested: the distribution of calls over the day is almost the same for the three Tuesdays, both normal and heavily-loaded. (Surprisingly, September 11 also catches up from around 13:00 or so.) For example, the hourly arrival rate during the peak hour—from 10:00–11:00—is about 2.5 times that of an average hour.

Instead of "Percent to mean", one can plot according to ***"Proportion to column totals"*** *which, in simple words, means the "hourly fraction-of-daily-load"*:

Going via the **"SEEStat"** tab, **"Output"**, **"Modify Tables and Charts"**, **"Proportions to column totals"**, and then **"OK".**

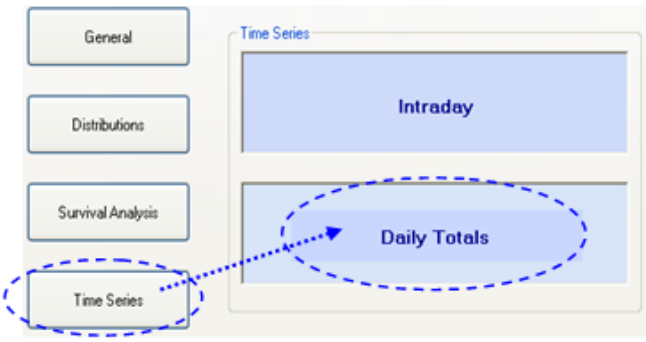


You see that the arrivals during Tuesday's peak hour 10:00–11:00 constitute about 10% of the daily total. (Significantly, such observations make load-predictions much easier: indeed, only the daily total must be predicted. Once the daily total is determined, the number of arrivals per hour is allocated according to the shape of the day; e.g. 10% allocated to 10:00–11:00.)

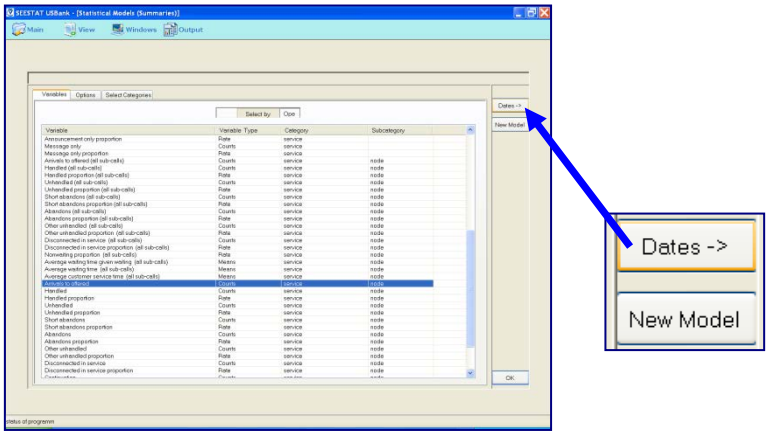
Example 1.3: *Time series (Daily totals)*

There are two types of daily-total time-series: individual days during a specific month and aggregated days by months. We now demonstrate these concepts.

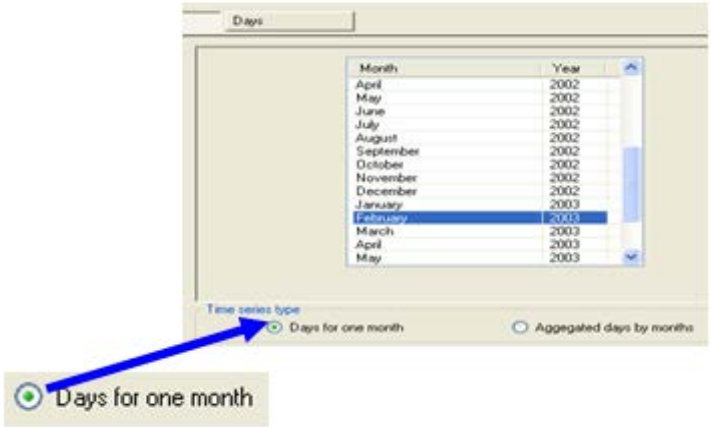
Click the **SEEStat** button on the task bar (left-bottom); next click **"Windows"** on the main menu (at the top) and select **"Statistical Models (Summaries)"**. Click the **"New Model"** button. Select **"Time Series"**, then **"Daily totals"**.



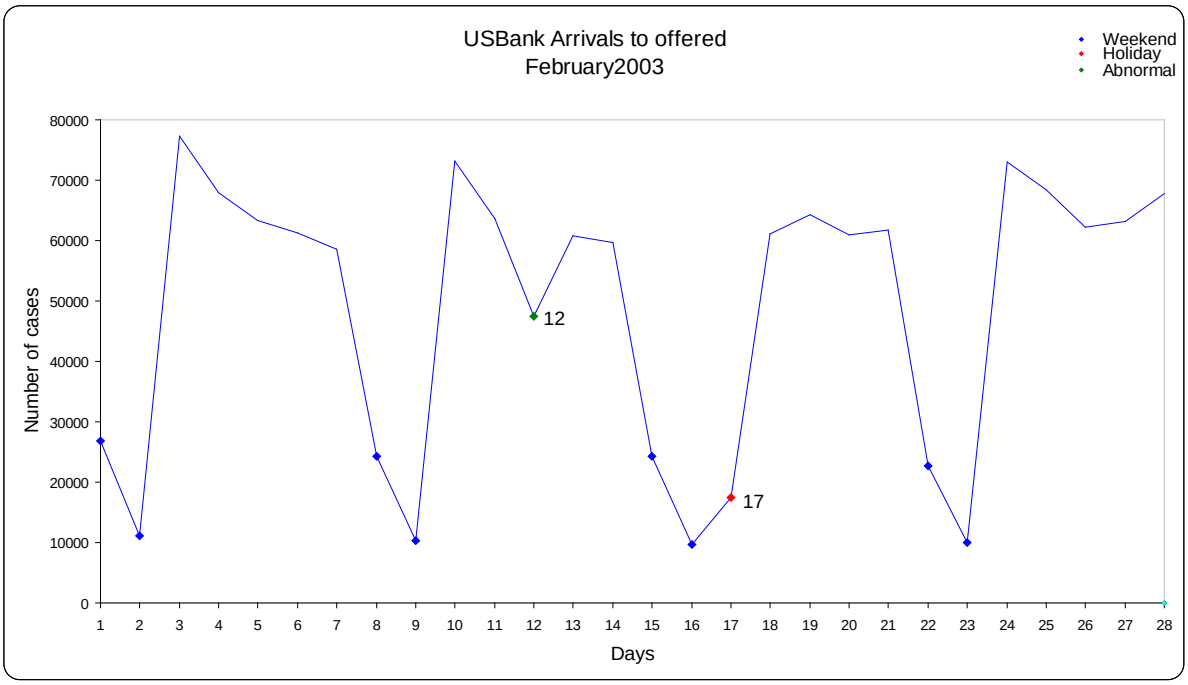
From the variables list select **"Arrivals to offered"** (around the middle of the list – it counts incoming calls that reached the queue for an agent service).
Click the **"Dates->"** button.



Mark **"Days for one month"** and select (after scrolling down) **February 2003.**

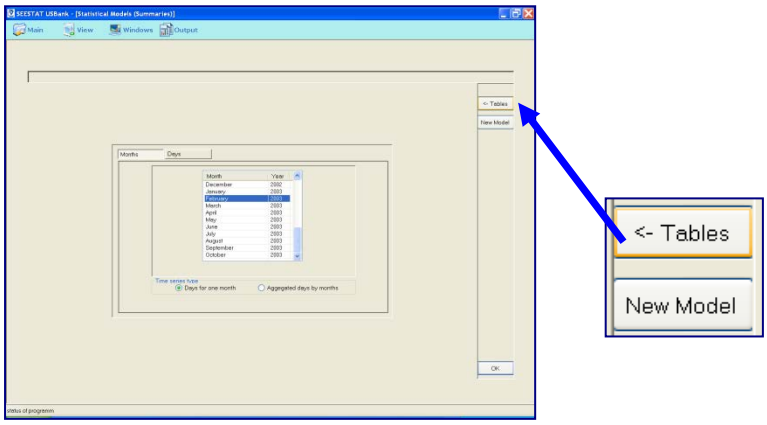


Open the **"Days"** tab; there is no need for you to select anything, but do note the Comments.
Click **"OK"**.



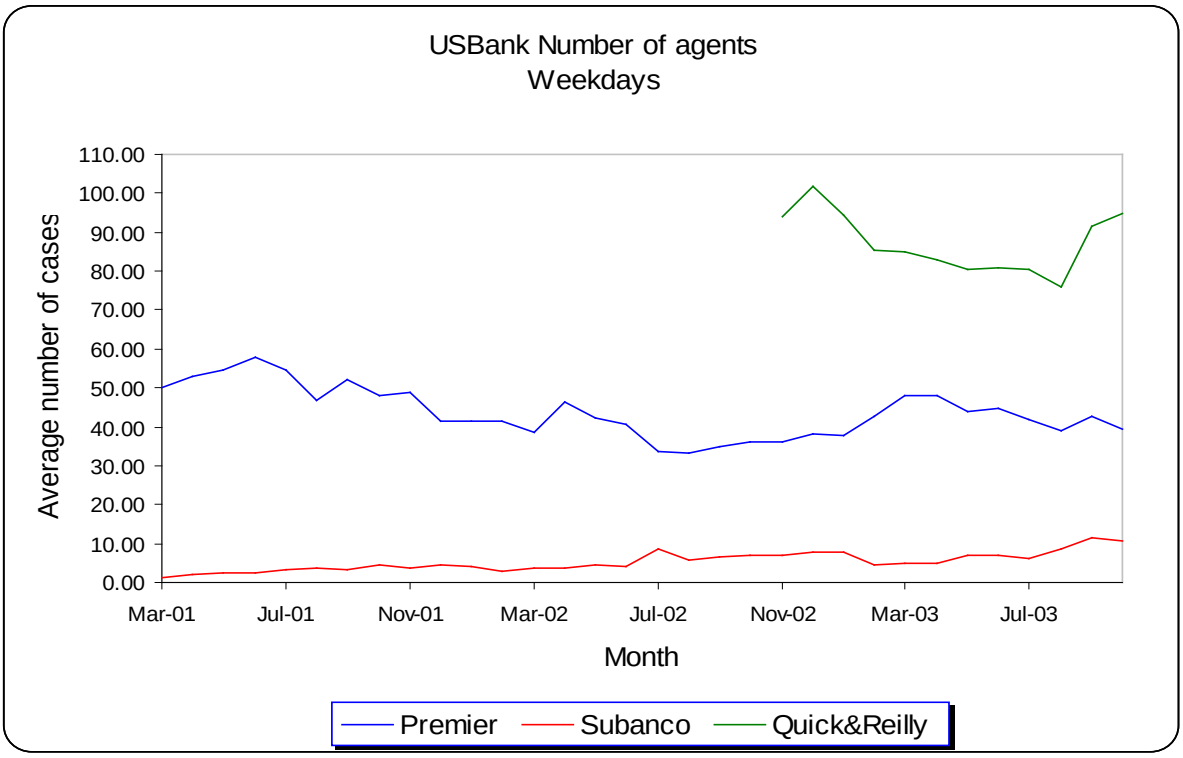
We first observe the weekly pattern in which weekdays have much higher arrivals than weekends (weekends are marked blue).
Now observe/recall that, on February 12 the system stopped working at 4:00 PM, and February 17 was a holiday—Washington's birthday. This is manifested on the chart, where these special days are marked as Abnormal (green) and Holiday (red).

Return to the **"Statistical Models (Summaries)"** window via the **SEEStat** tab.
(A reminder: Click the **SEEStat** button on the task bar (left-bottom), click **"Windows"** on the main menu (at the top) and select **"Statistical Models (Summaries)"**).
Click the **"<-Tables"** button (top right).



From the variables list select **"Number of agents"** (the first option).
Open the **"Select Categories"** tab. Select the following **three** services: **"Premier"** (priority Retail service) **"Subanco"** (Spanish language) and **"Quick&Reilly"** (brokerage). (In order to do so, hold down the **Ctrl** key and click the three options, one by one.)
Now click the **"Dates->"** button. Mark **"Aggregated days by months"** and click **"Select all"**.
Open the **"Days"** tab and select **"Weekdays"**.

Click **"OK"**.



You see that one of the selected services (Quick&Reilly) was integrated into the Call Center of USBank only in November 2002.

Exercise: Following the exact steps that you took above (to plot "Number of agents"), now plot the performance measure "Abandon proportion" (those who hang up after losing their patience).
Plot it per month over the whole period (from March 2001 till October 2003), for the four Categories of customers: Retail, Premier, Business, Platinum.

SEEStat will create 4 graphs: for simplicity, you can actually click on the graphs to identify the category and read off corresponding value.

Question: Would the managers of USBank be happy with what they see?
(Hint: compare % abandoning of Retail vs. Platinum or Business customers.)
Question: Can you explain the phenomenon you discovered? ("explain", not 'justify")?
(Hint: economy of scale – this is somewhat subtle and taught in a Service Engineering class.)

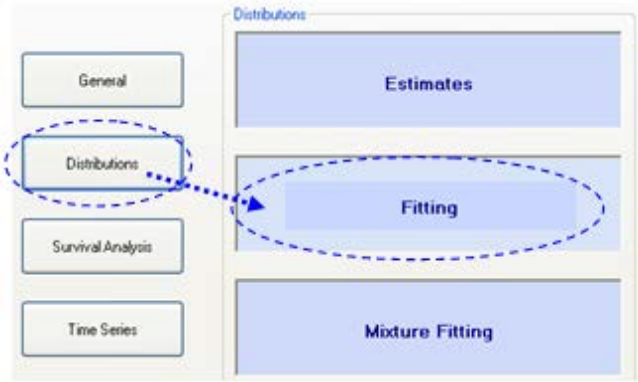
Part 2

Example 2.1: Distribution fitting

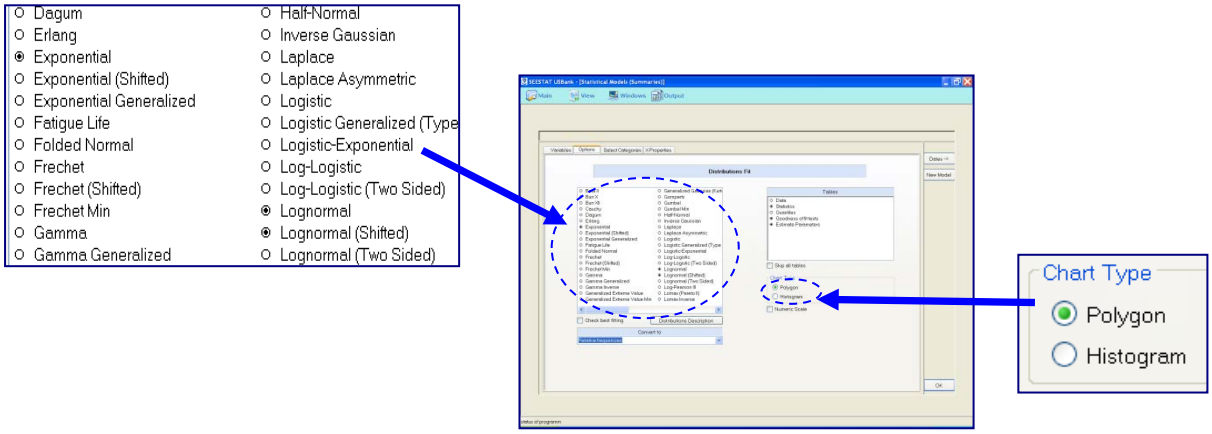
We now fit a parametric service-time distribution to the service-time data from Example 1.1

Open window **"Statistical Models (Summaries)"**. Click **"New Model"** and select

"Distributions" and **"Fitting"**.

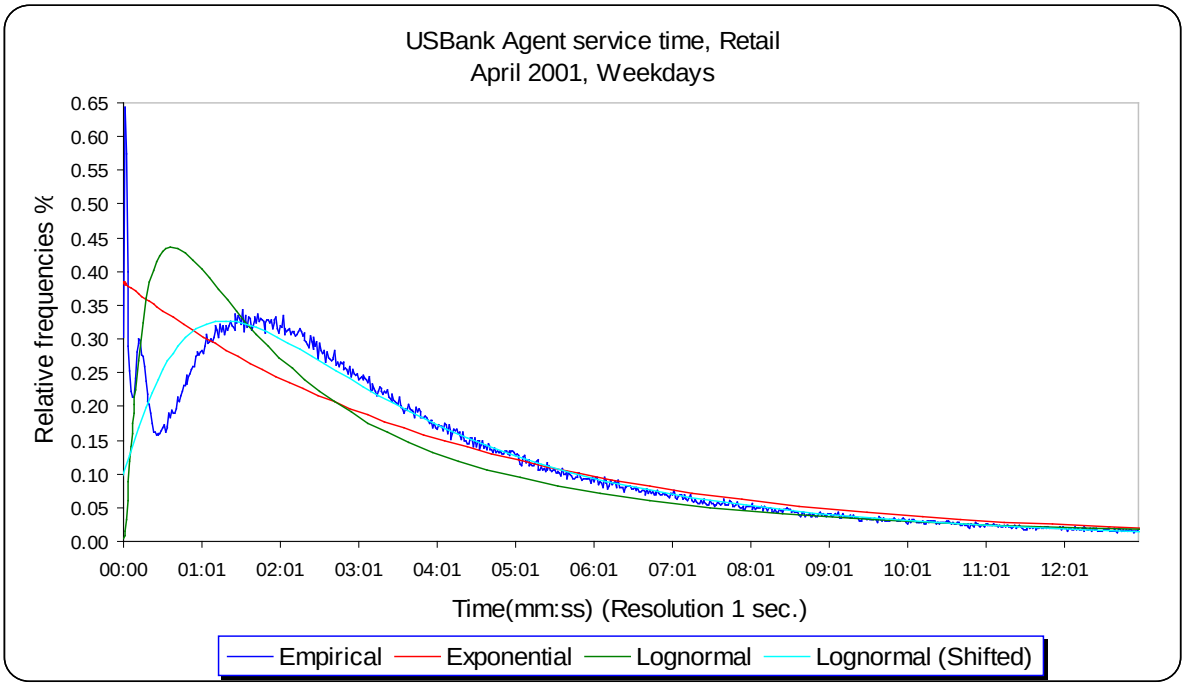


From the variables list select **"Agent service time"**.
Open the **"Options"** tab. You see the list of distributions available for fitting.
Mark simultaneously **3** of them: **Exponential, Lognormal,** and **Lognormal (Shifted)**.
Set chart type to **"Polygon"**. [1]



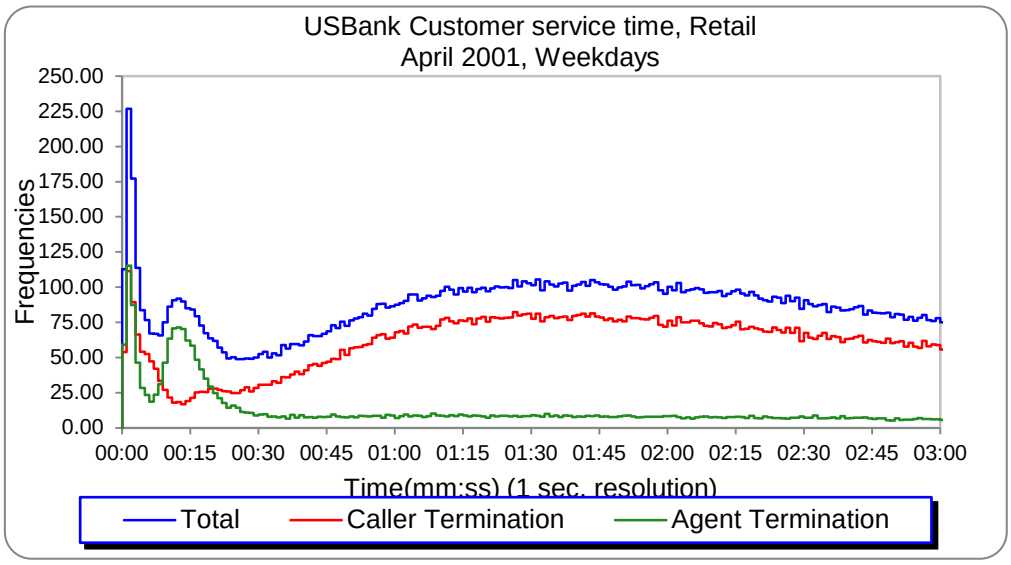
Open the **"X Properties"** tab and set resolution to **00:01** = 1 second.
Click the **"Dates->"** button. Select **April 2001** and **"Aggregated days"**, open the **"Days"** tab and select **"Weekdays"**.
Click the **"<-Tables"** button.
On the **"Select Categories"** tab select **Retail.**
Click **"OK"**

[1] SEEStat can automatically chose the best fit(s), out of 50 options, assuming that a fit exists which meets some accuracy-criteria.



Observe again the irregularities near the origin. It looks as though there are at least three distributions involved: very short calls, abnormally short calls (peak around 15 seconds) and, after around 30 seconds, the pattern looks rather regular. The best fit is produced by the Lognormal (Shifted) distribution (looks good after 1:30 min); but, clearly, close to the origin (from 00:00 till 1 minute) the fit is highly inadequate.

***Puzzle**: how can one explain the histogram-peaks of abnormally short phone calls (green graph)?*



***Note**: There is a difference between "Agent service time" and "Customer service time": the former covers more functions than the latter. Specifically:*
customer-service-time=talk-time+hold-time;
agent-service-time=talk-time+hold-time+wrap-up-time+call-type-time.
Here:

- *talk-time* – time the *caller* spent connected to a resource (agent, voice port, announcement, trunk, VRU), or duration that the *agent* spent connected to the caller
- *hold-time* – amount of time a caller spent on hold (an agent's teleset time) .
- *wrap-up-time* – amount of time an agent spent in a wrap-up state, after completion of the call segment (short times - about 1 second)
- *call-type-time* – amount of time an agent spent listening to a call type announcement prior to being connected to the call (whisper) (very short times - about 0.3 second on average)

You could use the tables on the previous sheet (the one accompanying the graph-sheet) to statistically validate the fit: ***scroll down*** *until reaching the Parameter-****"Estimates"*** *and* ***"Goodness-of-Fit Tests"*** *tables.*

Distribution	*Goodness-of-Fit Tests*				
	Residuals Std	*Kolmogorov-Smirnov*		*Cramer-von Mises*	
		Statistic	*p Value*	*Statistic*	*p Value*
Exponential	0.0333583	0.0648110	<.0001	688.91	<.0001
Lognormal	0.0504281	0.0878340	<.0001	1574.35	<.0001
Lognormal (Shifted)	0.0070425	0.0211673	<.0001	30.71	<.0001

Remark on EDA of large samples*: The above Table is only part of the full table in Excel. (There are also Anderson-Darling and Chi-Square tests – these do not change any of what will be now discussed.)*

The very small p-values clearly support a rejection of all 3 fittings above. However, our study is based on a very large sample size (over 620,000 observations), in which case essentially any null hypothesis would be rejected [2]. Thus, the concept of "statistical significance" loses its "practical significance". What does one then do?
Our starting point is the following general problem with traditional statistics: with a large enough sample, almost any difference or any correlation, say, will end up significant and lead to very small p-values. We then recommend the following 2 tests of statistical fit as well as visual fit. More precisely:

1. *Statistical fit: Check the standard-deviation of the residuals (referred to above as Residuals Std). This is a goodness-of-fit measure for how well the actual data points align with the theoretical model that we are trying to fit. A common practical threshold (I learned from researchers in my lab) calls for this standard deviation of residuals to be less than 2%.*
2. *Visual fit: there are formal methods to check visual fit (based on, for example, skewness, kurtosis and more). With very large sample sizes, an actual visual test in fact suffices.*

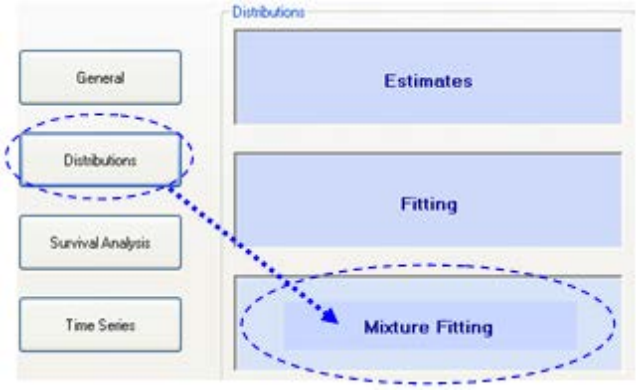
Consider, in our present example, the highlighted Lognormal (Shifted): the value of Residuals Std is indeed less than 2%, but the visual test reveals an excellent fit beyond 3 minutes, a reasonable fit between 1.5 and 3 minutes, and an unacceptable fit till 1.5 minutes. This suggests that the distribution is only "partially" LogNormal – a concept that is formally captured by fitting a mixture of distributions. We do this in our next example.

***Example 2.2*: Distribution mixture fitting**

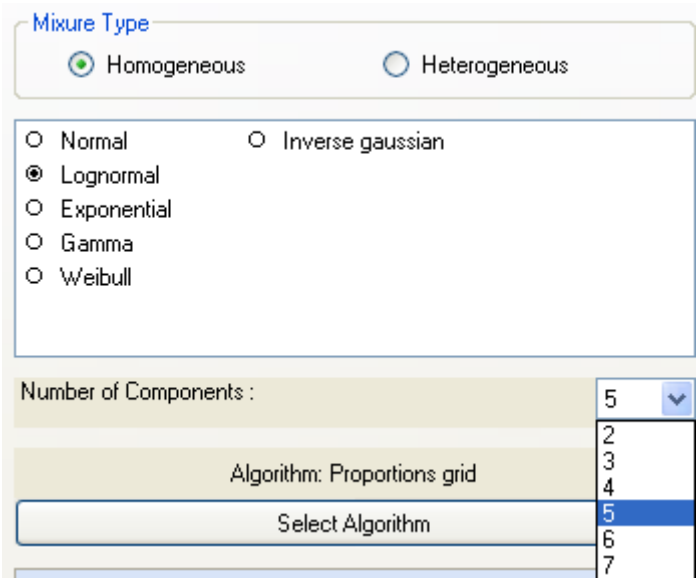
We now try to accommodate the behavior near the origin, of the "Agent service time" histogram in Example 2.1. We do that by a ***mixture*** *of distributions.*

[2] Suppose, for example, that "reality" is "exponential + 10^(-10)". Then large-enough-of-a-sample will detect this practically-insignificant deviation from exponentiality; in other words, the outcome of hypothesis testing will reject the null of exponentiality – it will be statistically-significant though practically-insignificant. In Tukey's words (1991, page 100), in the context of comparing group A with group B: "The effects of A and B are always different—in some decimal place—for any A and B." [Tukey J (1991) The philosophy of multiple comparisons. Statist. Sci. 6(1):100–116.]

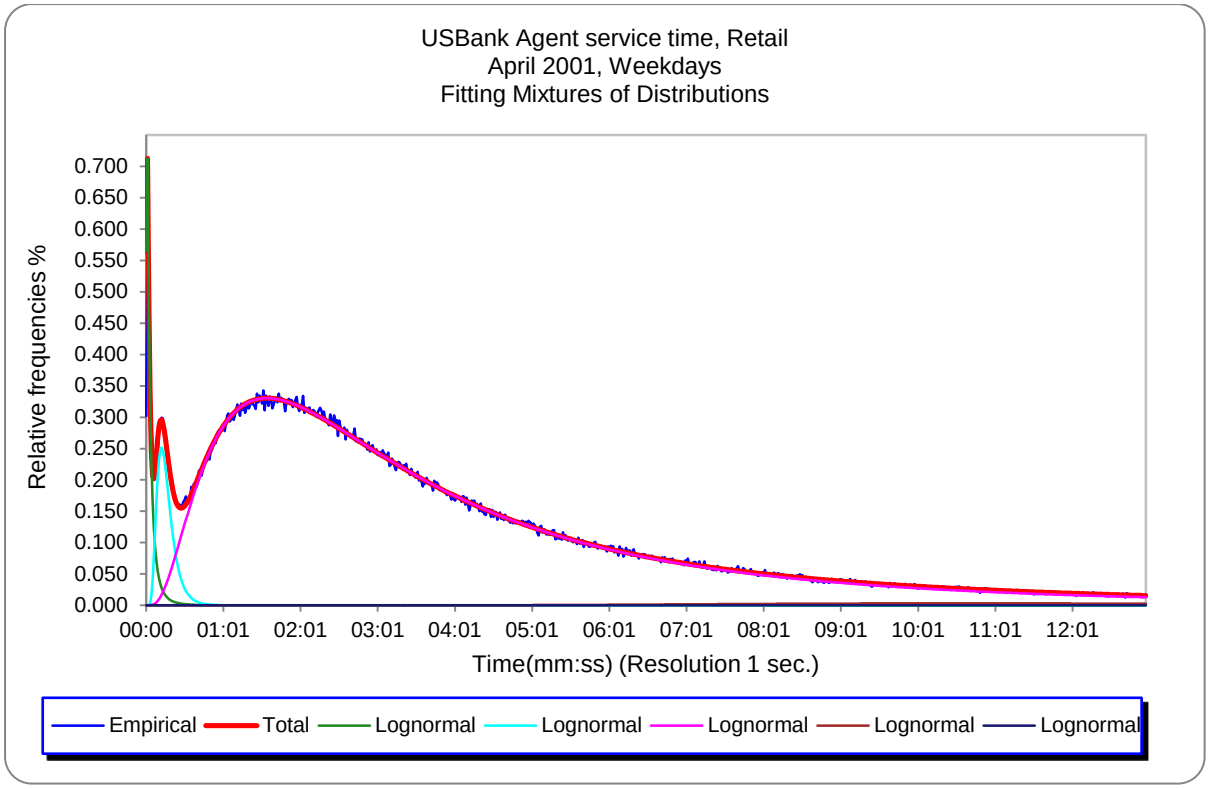
Via **SEEStat** return to the **"Statistical Models (Summaries)"** window, click **"New Model"**, select **"Distributions"** and **"Mixture fitting"**.



Open the **"Options"** tab. You can select a homogeneous or heterogeneous (mixture of various distributions) option. The former is the default. Select **"Lognormal"**. Set the number of mixture components to **5**, select chart type **Polygon**.



Click **"OK".**



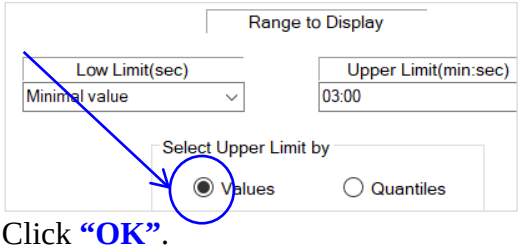
You observe an excellent fit (Red line) overall, which we verify and further analyze as follows:
Main component: Going to **"Table 2"** *in the Excel sheet, and scrolling down to the following Table (**Parameter Estimates**), one notes that the main component has a weight of 91% in the mixture—its role in the chart is to fit the part beyond 30 seconds, which it does very well.*

Parameter Estimates	
Components	*Mixing Proportions (%)*
1. Lognormal	3.19
2. Lognormal	3.55
3. Lognormal	91.09
4. Lognormal	1.83
5. Lognormal	0.34

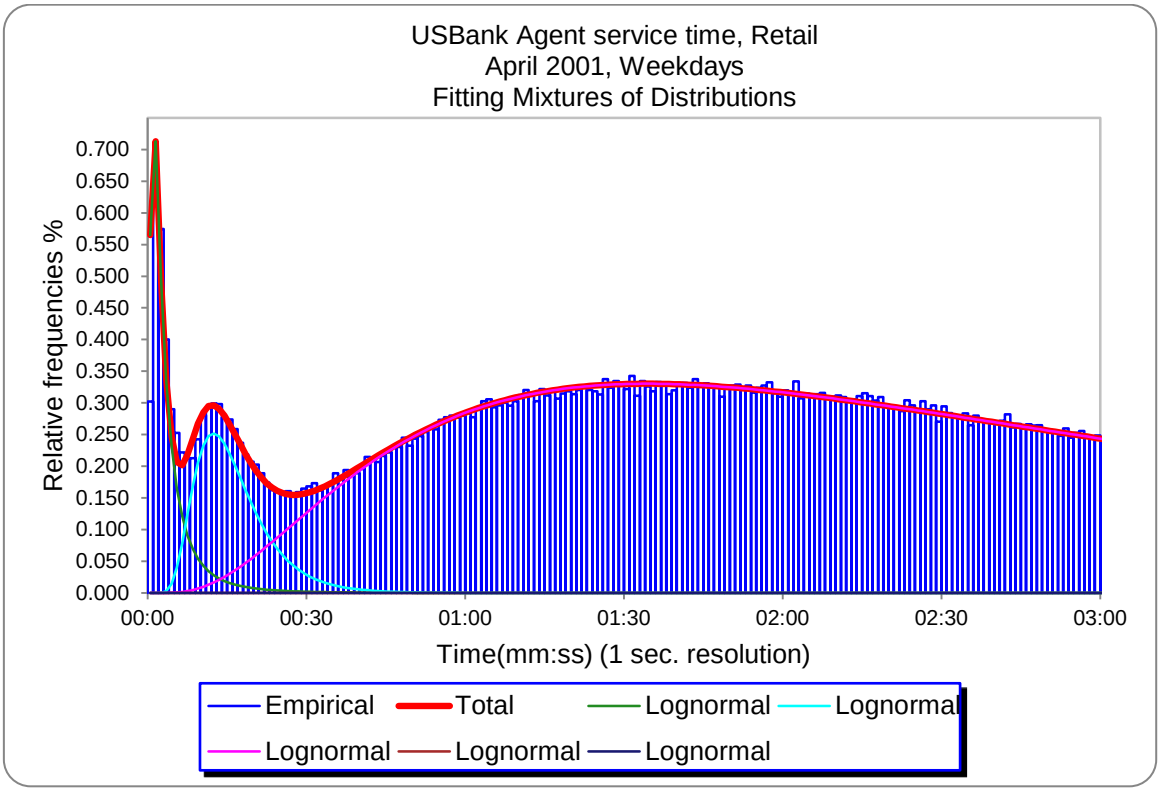
Small and large values, verified by zooming on small values and on large values: Starting with small, on the left side (near the origin), we discover 2 components that accommodate the very short and the abnormally short calls:

To zoom on the fitted components for small values (left side of distribution):
Click **"Output"**-> **"Modify Tables and Charts"**

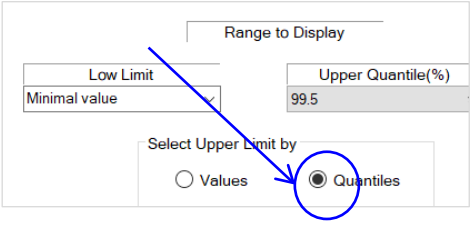
Open the **"Options"** tab. Select **"Histogram"**. Open **"Properties"** tab. Select **"Values"** and upper limit 03:00 (3 minutes).



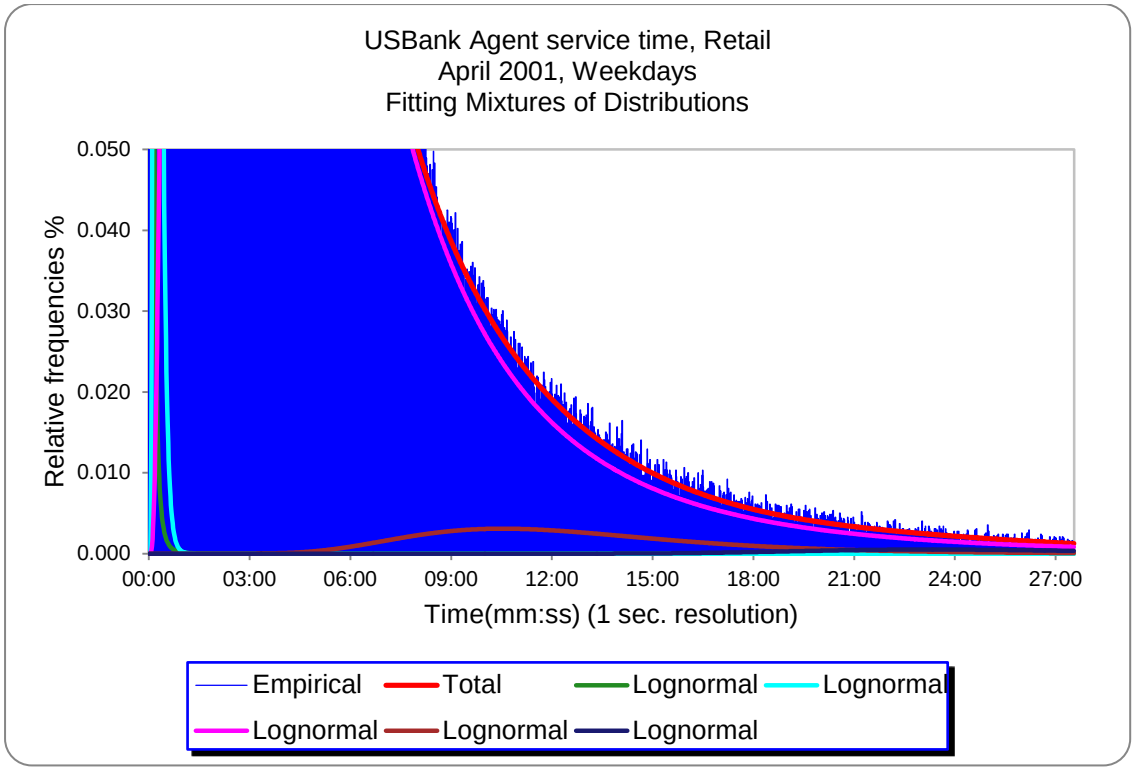
Click **"OK"**.



To zoom on the fitted components for large values (right tail of distribution):
Click **"Output"**-> **"Modify Tables and Charts". "**. Open **"Properties"** tab. Select **"Quantiles"** and upper quantile 99.5.



Click **"OK"**. Right click on horizontal axis->format axis->axis option-> maximum->0.05.



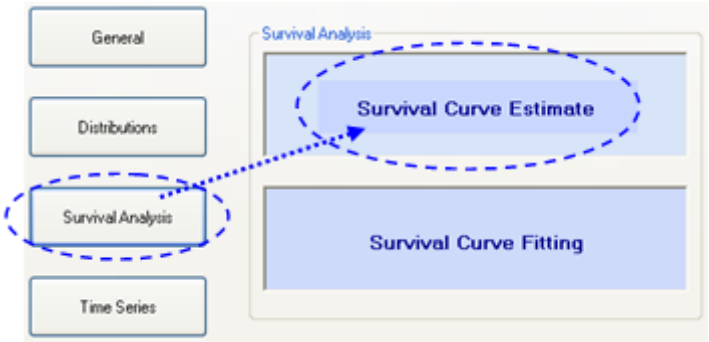
See Appendix A for a short explanation on Mixture-Fitting and the algorithms used for distribution fitting.

Example 2.3: ***Survival analysis with smoothing of hazard rates***

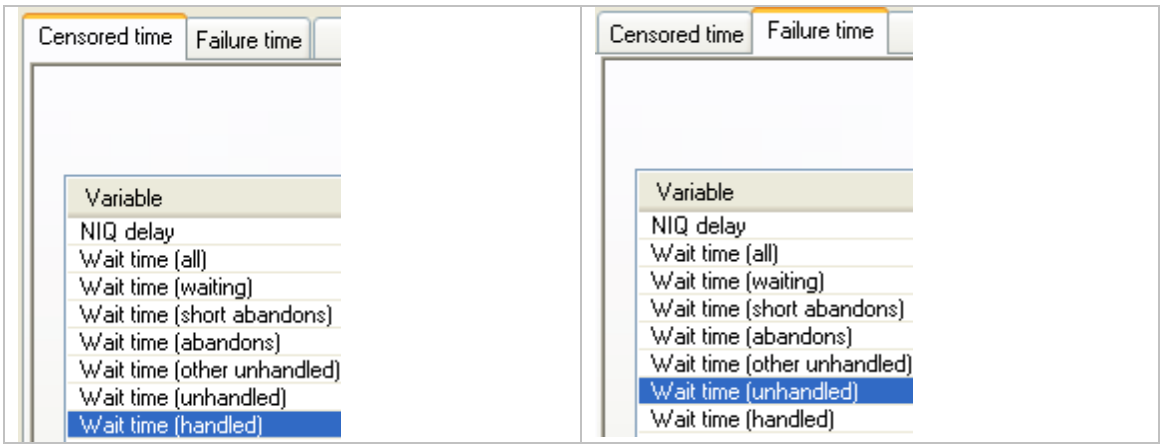
SEEStat supports several survival models. These are required, for example, in order to analyze the ***(im)patience*** *of customers, namely the time customers are willing to wait until hanging up (due to impatience). Indeed, for those customers who got served, their waiting-time in queue provides only a lower bound on how long they are willing to wait—their (im)patience data hence constitutes what one refers to as (right-)****censored*** *observations. One must thus "uncensor" the data to produce adequate estimates of (im)patience.*
To this end, we now use some tools from Survival Analysis: these will produce hazard-rate functions, which provide natural statistical summaries of (im)patience.

Return, via **SEEStat**, to the **"Statistical Models (Summaries)"** window, click **"New Model"**.

Select **"Survival analysis"** and **"Survival Curve Estimate"**.



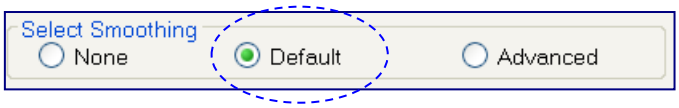
There are two variable tabs. The first tab **"Censored time"** is open. Select **"Wait time (handled)"**: this corresponds to the waiting time of the customers who received service (handled). Open the **"Failure time"** tab and select **"Wait time (unhandled)":** this corresponds to the waiting of customers who joined the queue but did not receive service (mainly due to abandonment, though there are sometimes other reasons such as system malfunction).



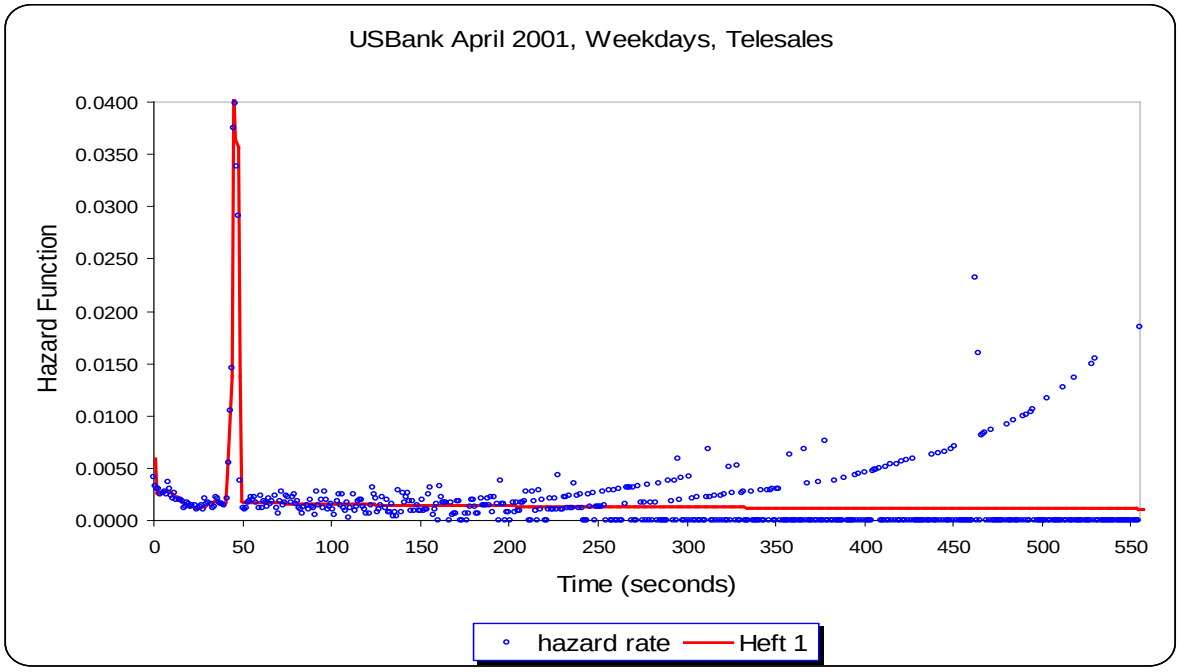
We are now going to estimate (the hazard-rate of) the **time-to-Failure**, which here stands for **time-to-Abandon**, or **(im)patience** as referred to above.
Note that the waiting-time of a customer who was served/handled provides only a lower-bound on that customer's (im)patience – we refer above to that waiting-time as **Censored time**. (**Wait time (handled)** is the time to failure that has been censored to the observed value – therefore it is referred to as **Censored time**.)

Back to our "journey":

Open the **"Options"** tab. SEEStat supports several methods of smoothing, which are applicable to hazard rates and beyond. Select **"Default"** smoothing (which, this time, happens to be the method of HEFT).



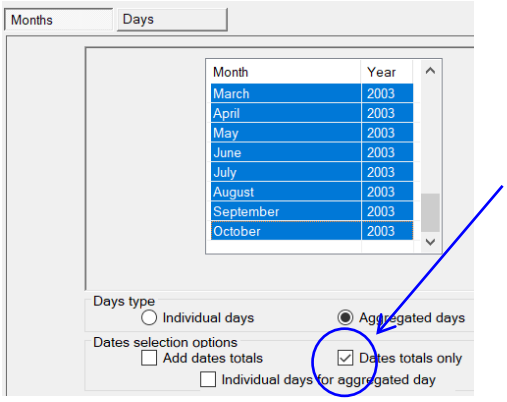
From the tab **"Select categories"** select **"Telesales"**.
Click **"Dates"**. Select **"April 2001"** and on the tab **"Days"** select **"Weekdays"**.
Click **"OK".**



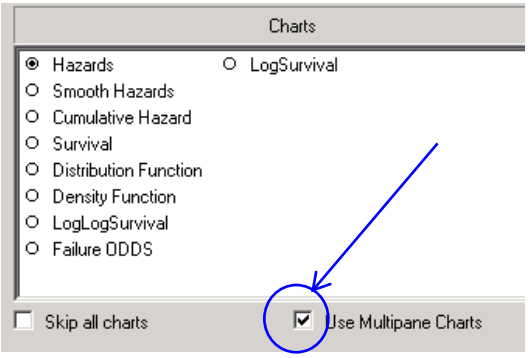
A noticeable peak in the hazard-rate indicates that there is a trigger for customers to abandon after about 50 seconds of waiting (which, based on our experience, is likely to be the result of a voice-announcement at that time: such announcements, regardless of their content, "reminds" customers of their wait and thus increase their likelihood of abandonment).

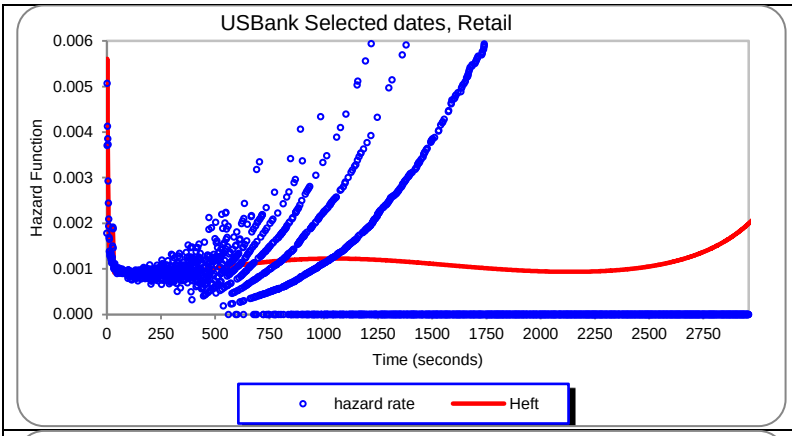
Remark: There are ample smoothing methods around. Choosing the right/best – a process with an art-component in it – calls for a tradeoff between over-smoothing (losing details) and under-smoothing (leaving too much noise).
See Appendix C for References to some of the smoothing algorithms that SEEStat uses.

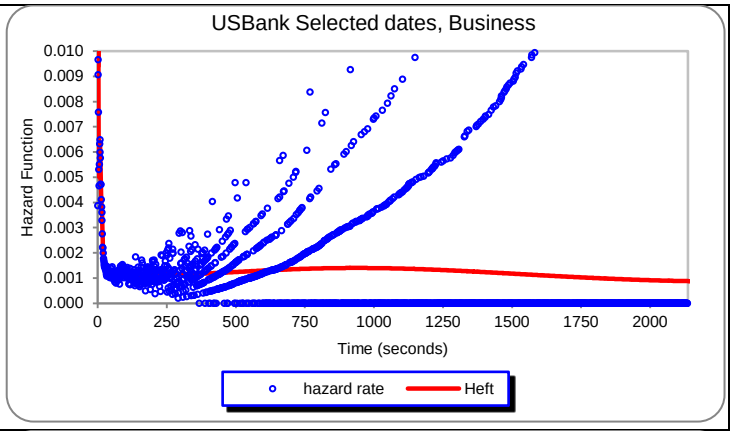
Hazard rates of (im)Patience takes, in practice, many shapes and forms. We now create some of them simultaneously.
Return, via **SEEStat**, to **"Statistical Models (Summaries)".** Select **"Dates totals only",** then select all months from March 2001 to October 2003. Click **"Days"** and **"All days".**

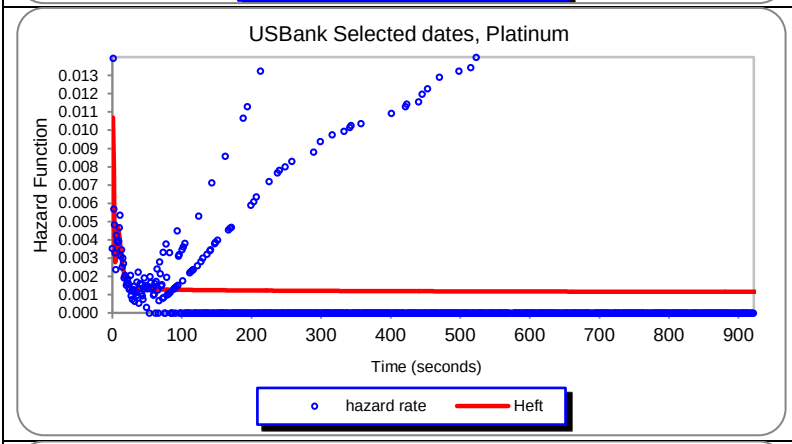


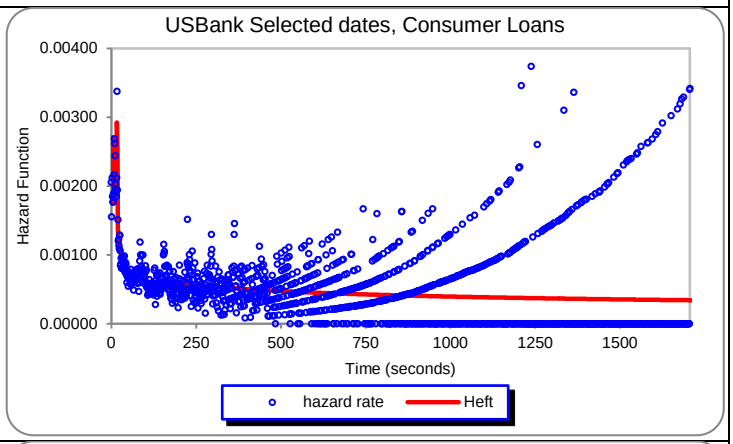
Click **“<-Tables”.** Open the **"Selects Categories"** tab and select all categories *except* for *Total* and *Premier*. Open the **"Options"** tab. Select **"Use Multipane Charts".** Click **"OK".**

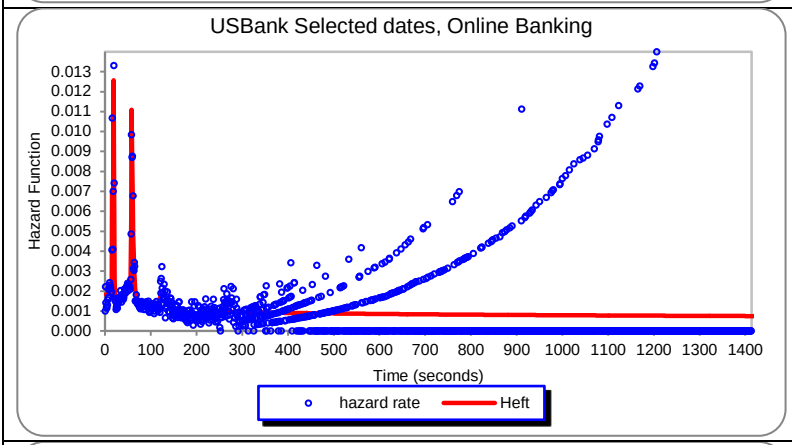


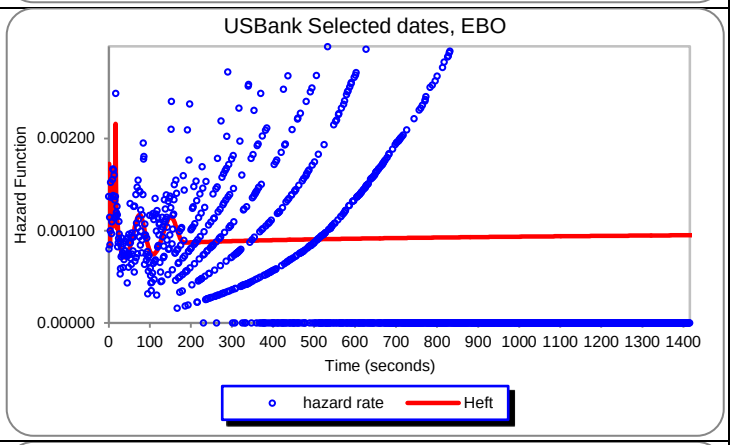

USBank Selected dates, Retail
Hazard Function
Time (seconds)
hazard rate
Heft

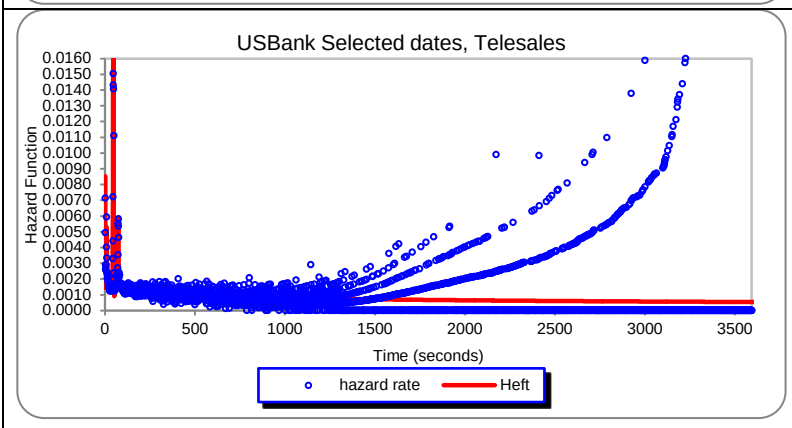

USBank Selected dates, Business
Hazard Function
Time (seconds)
hazard rate
Heft

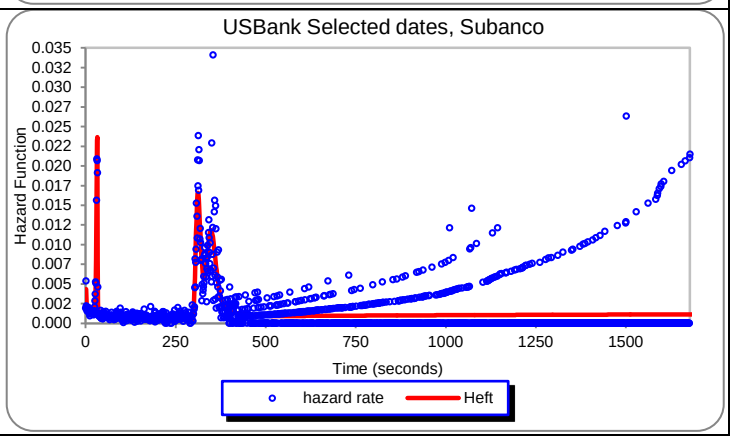

USBank Selected dates, Platinum
Hazard Function
Time (seconds)
hazard rate
Heft

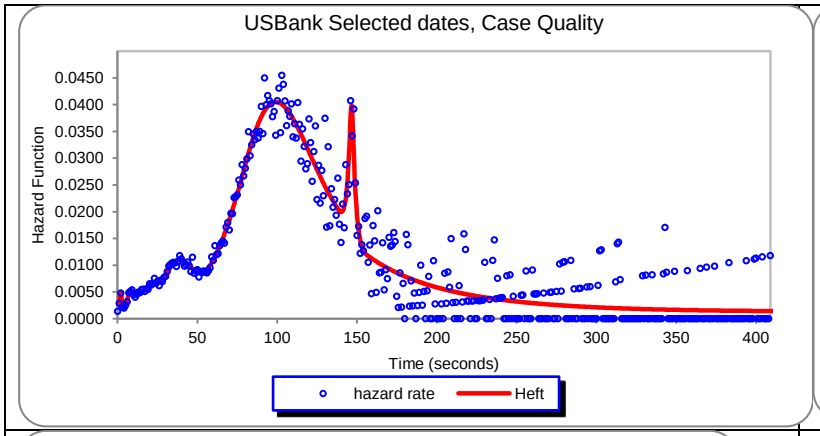

USBank Selected dates, Consumer Loans
Hazard Function
Time (seconds)
hazard rate
Heft

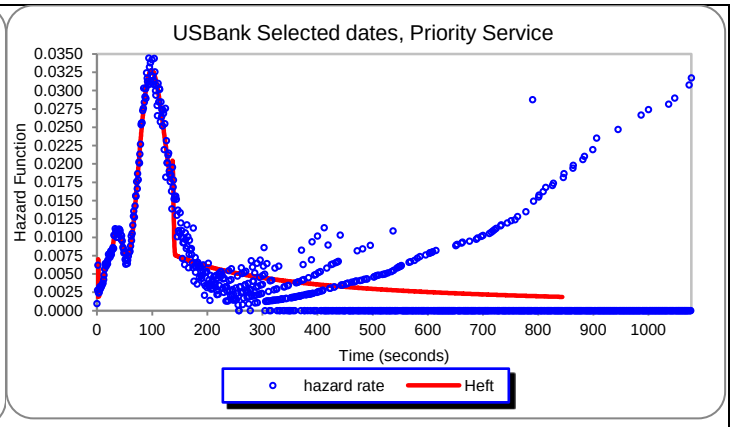

USBank Selected dates, Online Banking
Hazard Function
Time (seconds)
hazard rate
Heft

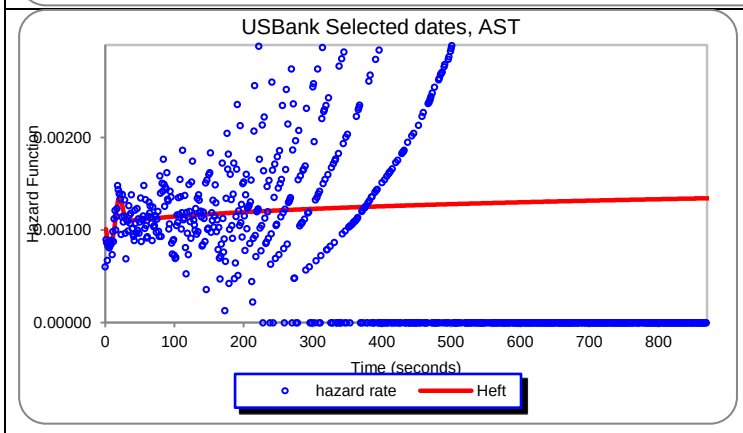

USBank Selected dates, EBO
Hazard Function
Time (seconds)
hazard rate
Heft

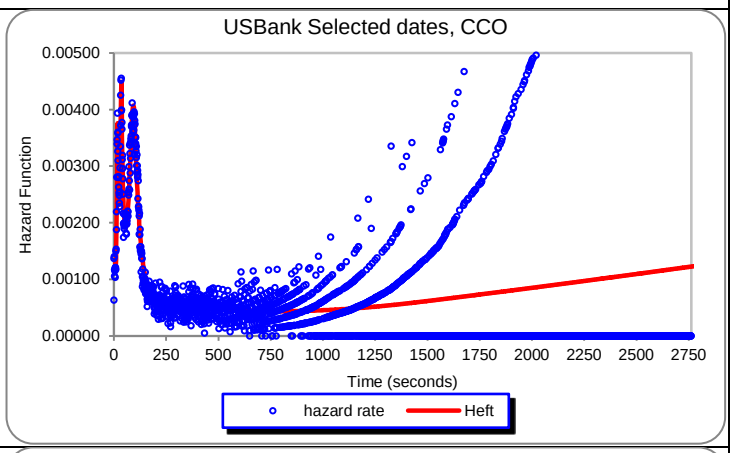

USBank Selected dates, Telesales
Hazard Function
Time (seconds)
hazard rate
Heft

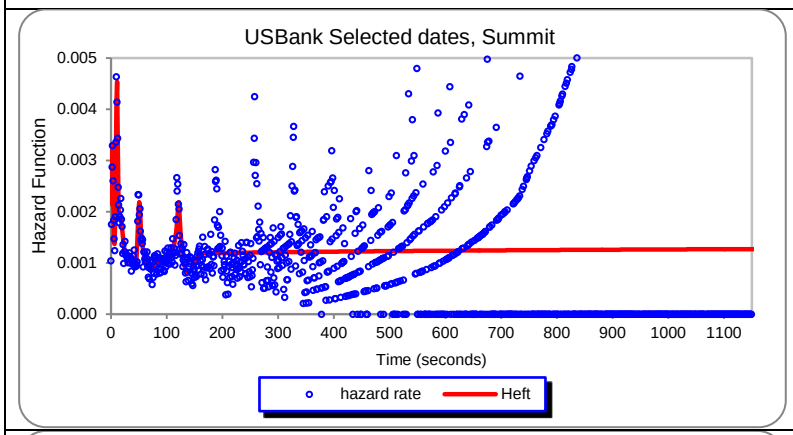

USBank Selected dates, Subanco
Hazard Function
Time (seconds)
hazard rate
Heft

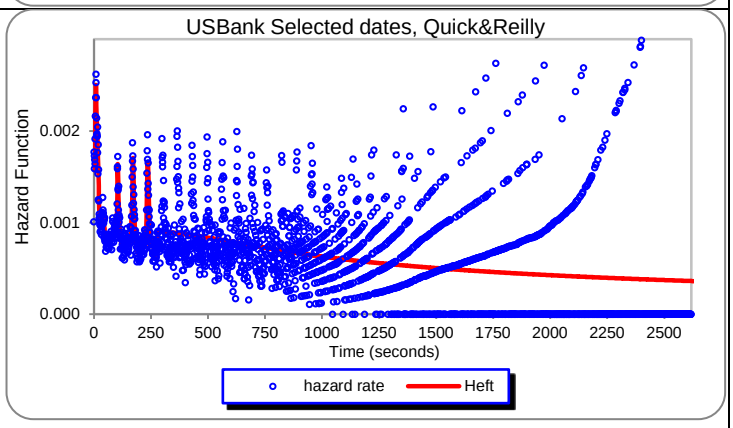

USBank Selected dates, Case Quality
Hazard Function
Time (seconds)
hazard rate
Heft

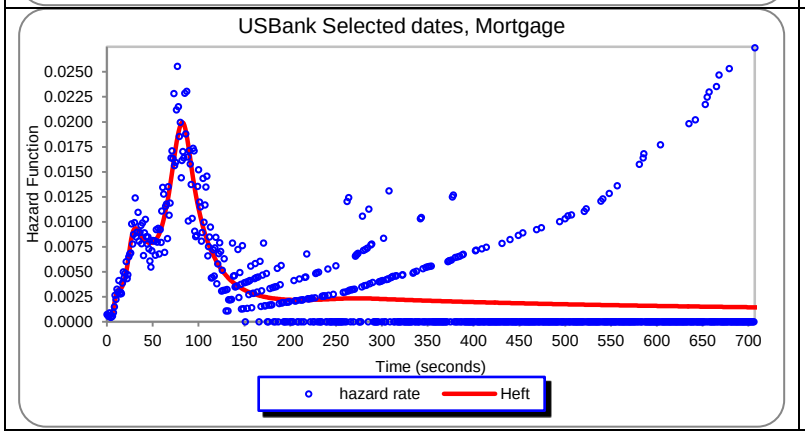

USBank Selected dates, Priority Service
Hazard Function
Time (seconds)
hazard rate
Heft

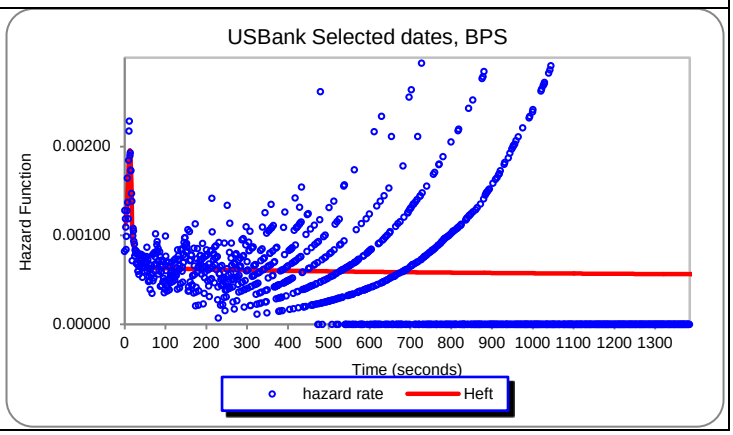

USBank Selected dates, AST
Hazard Function
Time (seconds)
hazard rate
Heft


USBank Selected dates, CCO
Hazard Function
Time (seconds)
hazard rate
Heft


USBank Selected dates, Summit
Hazard Function
Time (seconds)
hazard rate
Heft


USBank Selected dates, Quick&Reilly
Hazard Function
Time (seconds)
hazard rate
Heft


USBank Selected dates, Mortgage
Hazard Function
Time (seconds)
hazard rate
Heft


USBank Selected dates, BPS
Hazard Function
Time (seconds)
hazard rate
Heft

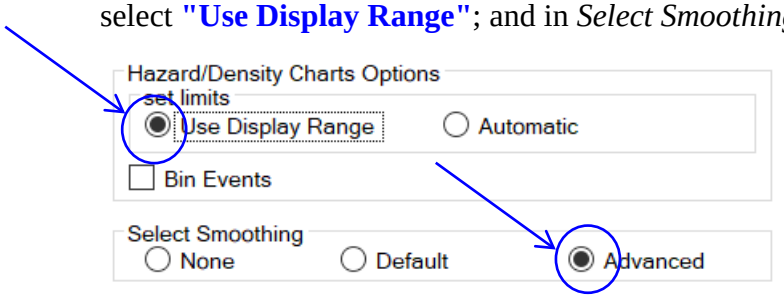
***Puzzle**: why do most of the above graphs end with a similar pattern – somewhat parallel convex increasing "lines"? The reason is the algorithm of "un-censoring", or more specifically how the estimator's formula behaves under small samples. Indeed, the estimator at time t is calculated as follows:*

h(t) = {number abandoning during period [t, t+1)} / {number still waiting at time t}.

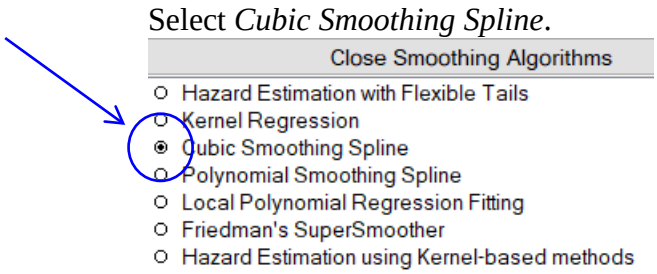
The data points h(0), h(1), h(2),…, are displayed as blue circles. Applying a smoothing algorithm (several are available), SEEStat creates and displays a continuous-time function h(t), t >=0 (in red).
Now consider all the time-periods where the number of abandoning is small, say 0, 1, 2, 3, 4. (These periods become more frequent as time t gets large.) The value 0 generates the circles on the x-axis; the value 1 creates the shape of 1/x, then 2/x etc.

Making process-sense (in addition to statistical-sense) of the above "shapes and forms" requires further analysis. We demonstrate this via the example of "Quick&Reilly" service.

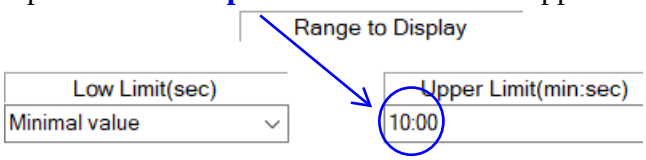
Return, via **SEEStat**, to the **"Statistical Models (Summaries)"**. In the tab **"Options"** select **"Use Display Range"**; and in *Select Smoothing* select **"Advanced".**



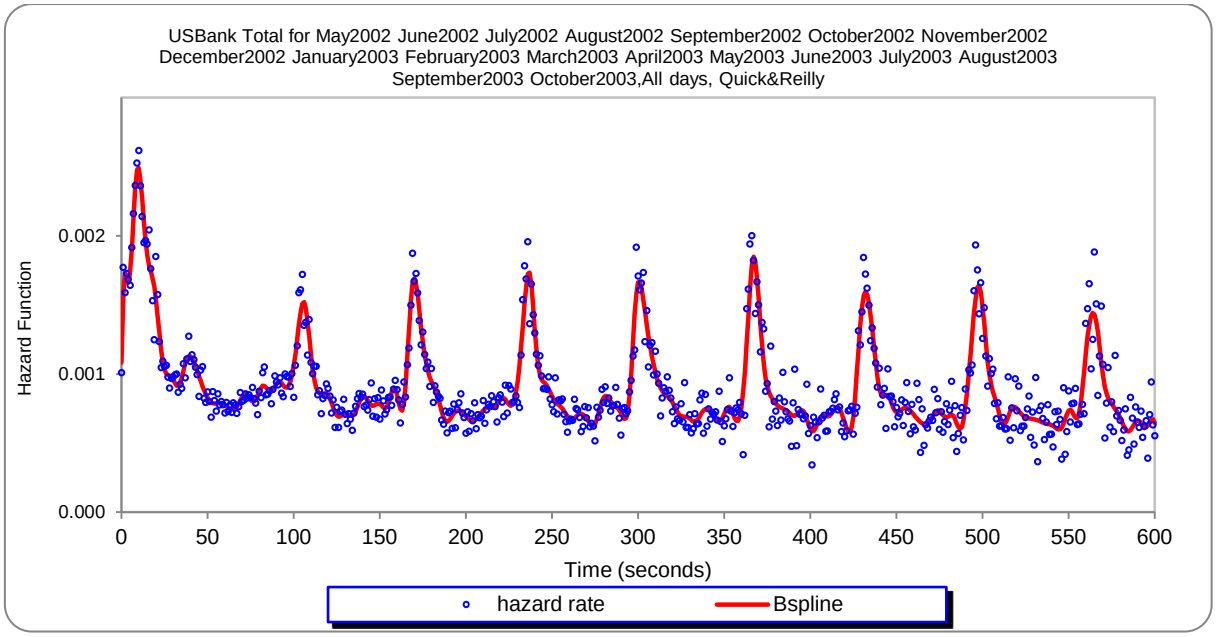
Select *Cubic Smoothing Spline*.



Open the **"X Properties"** tab and fill in upper limit 10:00 (10 minutes).



Open the **"Select Categories"** tab and select **"Quick&Reilly"**. Click **"OK".**

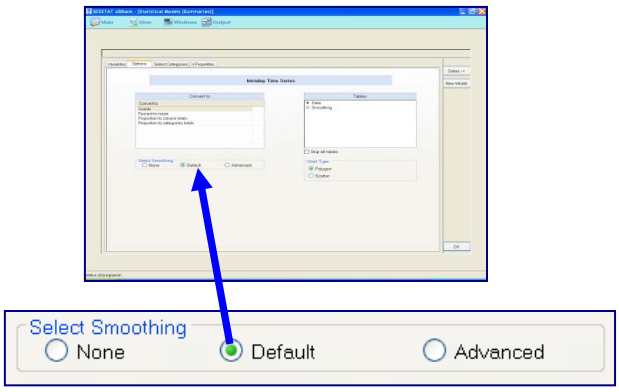


I hope that you agree that the above graph is nothing but beautiful.
***Puzzle**: The above function is an "uncensored" hazard-rate of the time that a customer is willing to wait (customers' patience). Why do peaks occur every minute approximately? We shall return to this Puzzle in Example 3.1 below.*

Example 2.4: ***Smoothing of intraday time series***

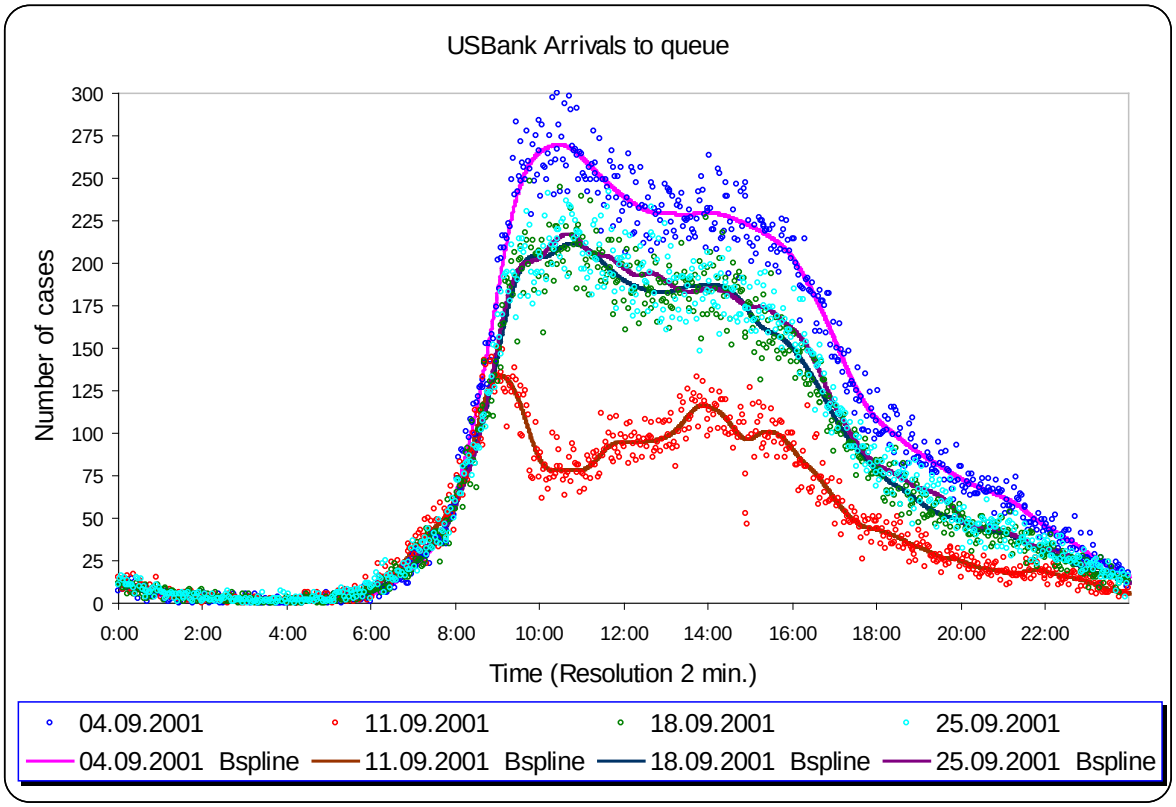
Smoothing algorithms are available for several statistical models (e.g. for hazard-rates, as in the previous Example 2.3; See Appendix C). We now demonstrate the application of smoothing to the data used in Example 1.2.

Return as usual to **"Statistical Models (Summaries)"**, click **"New Model"**, select **"Time Series"** and **"Intraday"**. Select **"Arrivals to queue"**. In **"Options"** tab select **"Default"** smoothing. (This time, the default method is Cubic Splines, which is referred to as Bspline in graphs).



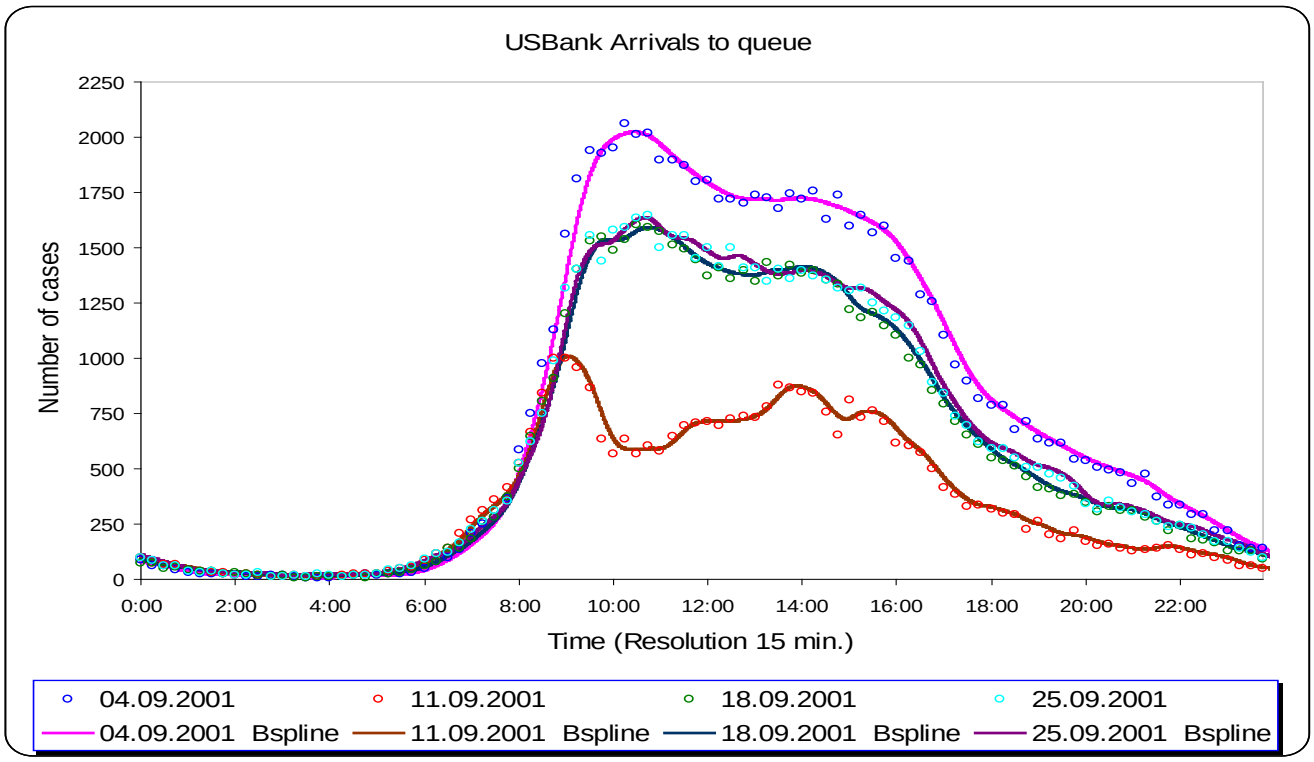
Select **"Scatter"** as Chart Type.
In the **"X Properties"** tab, set resolution to **02:00** = 2 minutes.
Click **"Dates"**, mark **"Individual days"** as Days type, and select **"September 2001"**.
In the **"Days"** tab, if not already selected, select (with **"Ctrl"** and click) all four **Tuesdays** of September.

Click **"OK"**



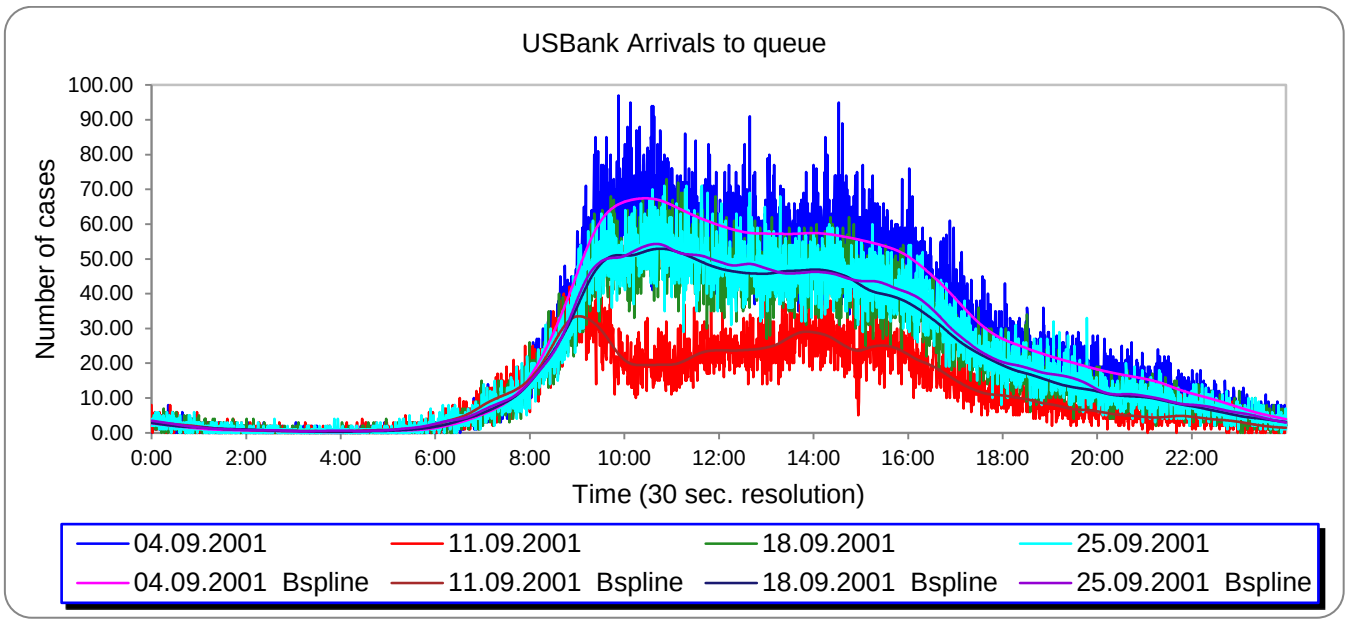
For this small resolution of 2 minutes, there is plenty of noise, but the smoothed data clearly identifies the regular pattern that was discovered before. (Note that the smoothed curves are computed with the minimal resolution for this variable, which is 30 seconds as can be seen in **"X Properties"***; the 2-minute resolution is only for display.)*

Click **"Output"** on the main menu, then click **"Modify Tables and Charts"**.
Open the **"Properties"** tab, set resolution to **15 min** and click **"OK"**.



The Averaged Data (over 15 minutes) is now much closer to the smoothed curves, and the Bspline curves are unchanged – both as expected.

Finally, repeat the above, via **"Modify Tables and Charts"**, but now with resolution of **00:30** seconds (the highest), and with selecting Chart Type **"Polygon"** (rather than "Scatter"). *The erratic "noise band" around each smooth curves represents the uncertainly-challenge that call-center managers must address, continuously during each day. (00:01 sec resolution is even more just.)*

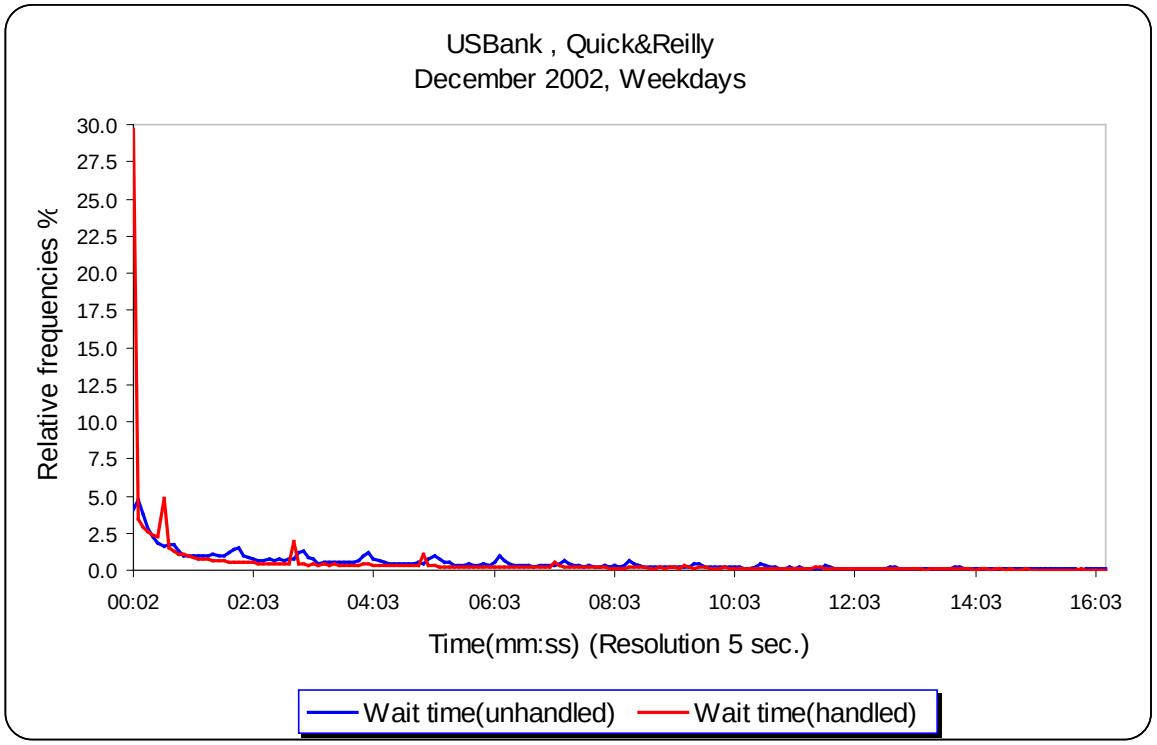


Part 3

Example 3.1: *Queue regulated by a protocol & announcements*

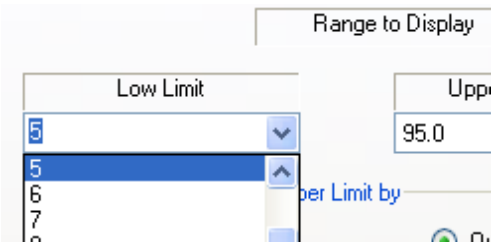
Via **SEEStat** return to the **"Statistical Models (Summaries)"** window, click **"New Model"**, then click the **"Distributions"** button. Three available distribution models appear. Select **"Estimates"**. In the **"Variables"** tab select (using **Ctrl**) both **"Wait time (unhandled)"** and **"Wait time (handled)".**

In the **"Options"** tab select Chart Type **Polygon**. Click "**Dates->"**, select **December 2002**, make sure the **"Aggregated days"** option is selected, and in **"Days"** select **Weekdays**. Click **"Tables".** In **"Select Categories"** select **"Quick&Reilly".** Click the **"X Properties"** button**,** and select **"Upper Quantile (%)"** to be **95%** (if it is not already such). Click "**OK**".

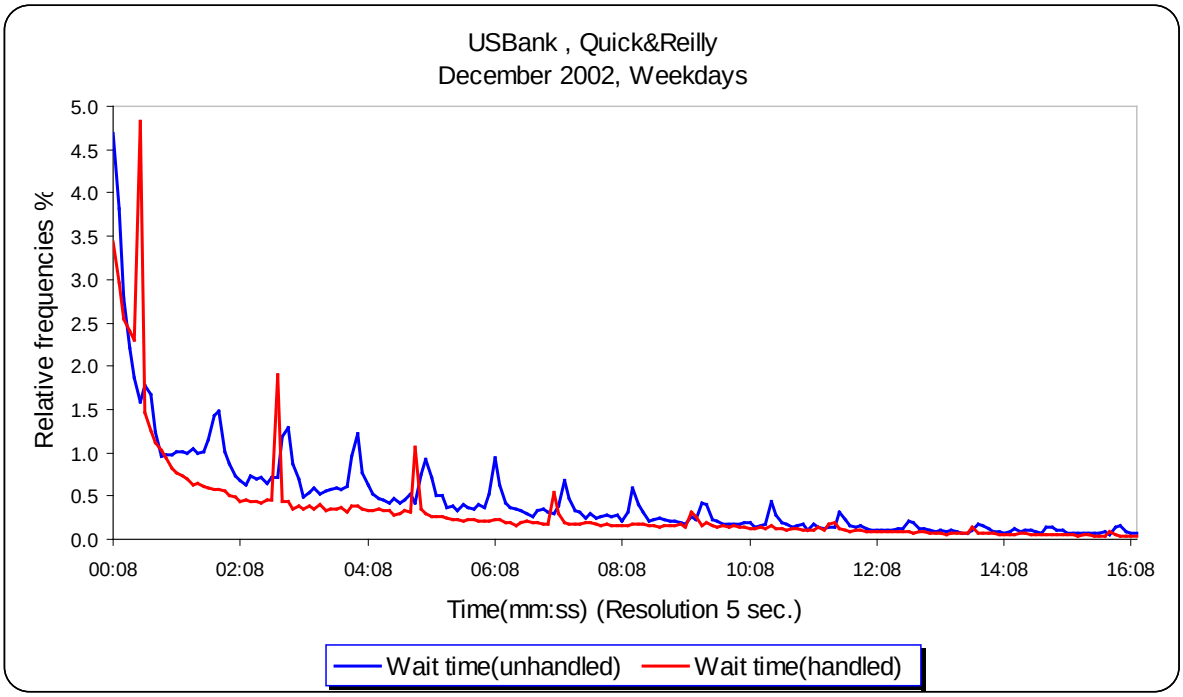


Looks like both graphs are periodical, but details are too small for comfort. To get a better focus, you now cut the chart on the left side.

Click **"Output"** on the main menu and then **"Modify Tables and Charts"**.
Open **"Properties"**, set the **low limit** to **5 seconds**.



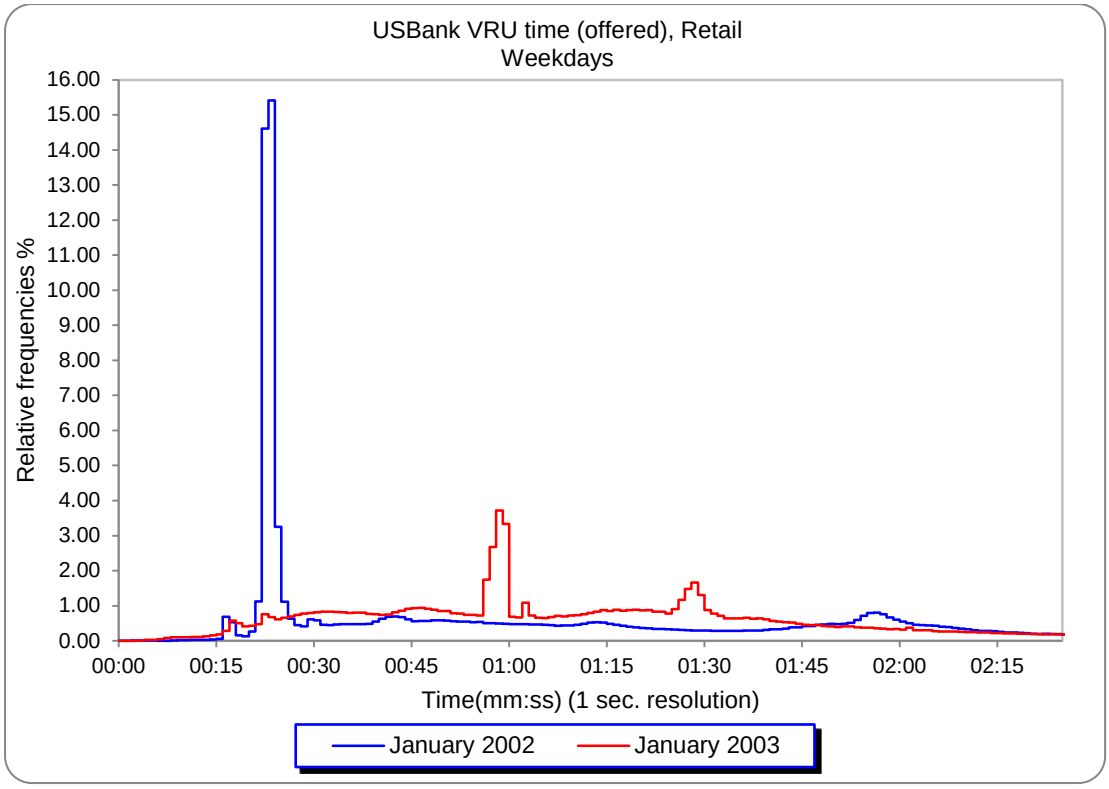
Click **"OK"**.



As you see, the Wait time (unhandled), in blue, peaks every 65 sec. The Wait time (handled), in red, peaks every 130 seconds. These interesting observations are yet to find their founded explanations, but our experience suggests the following explanations: peaks in the "Wait-time (unhandled)" are "psychological", for example a reaction of a customer to an announcement (here every 65 seconds); and peaks in the "Wait-time (handled)" are "protocol-driven", for example a result of a priority upgrade (here every 130 seconds.)

Example 3.2: *VRU-time is protocol-driven, BUT the protocol changes in time*

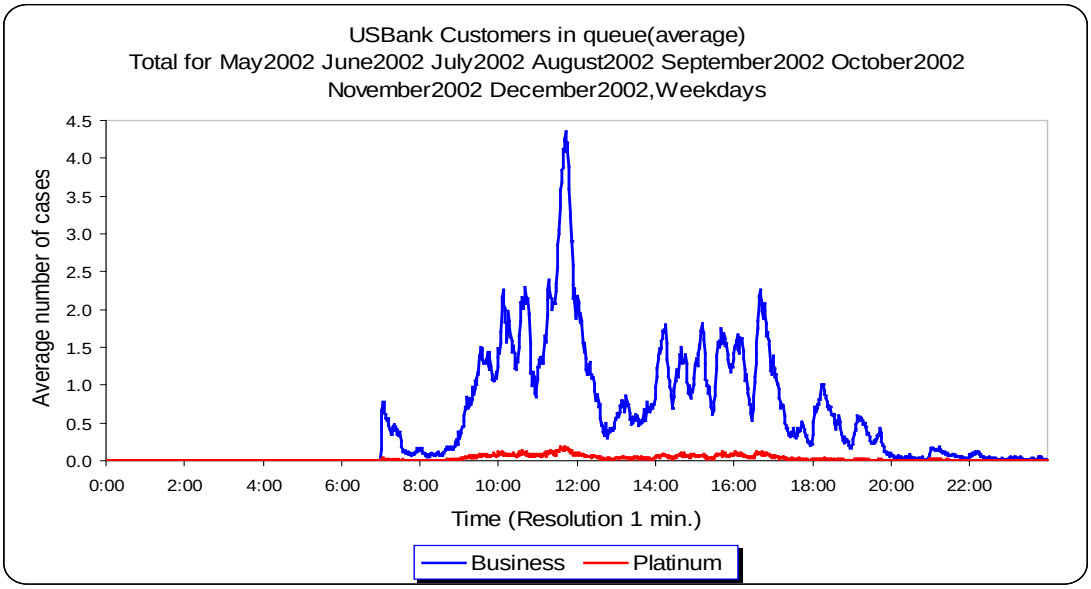
Click **"Main"->"Statistical Models (Summaries)"**.
Click the **"Distributions"** button. Three available distribution models appear.
Select **"Estimates".** In the **"Variables"** tab select **"VRU time (offered)".** In the **"Options"** tab select chart type **Histogram**. In **"Select Categories"** select **"Retail".**
Open **"Properties"**, select resolution **00:01** (1 second), and change **upper quantile** limit to **90.0**.
Click **"Dates->"**, select (using **Ctrl**) both **"January 2002"** and **"January 2003".** Click **"Days"** and **"Weekdays".**
Click "**OK**".



This is an example of where the VRU protocol, for customers who opted to be served by an agent, was changed (for a reason unbeknown to us). Specifically, VRU time (offered) - the total time spent in the VRU, for customers who then enter the agents queue) – peaks in January 2002 (blue) at 22 seconds (tall peak) and before 2 minutes (low peak). In January 2003 (red), on the other hand, there are peaks at 58 seconds and 1.5 minute. Such change-of-peaks results from a <u>change in the protocol</u> that supports/manages customer transfers from the VRU to the agents queue.

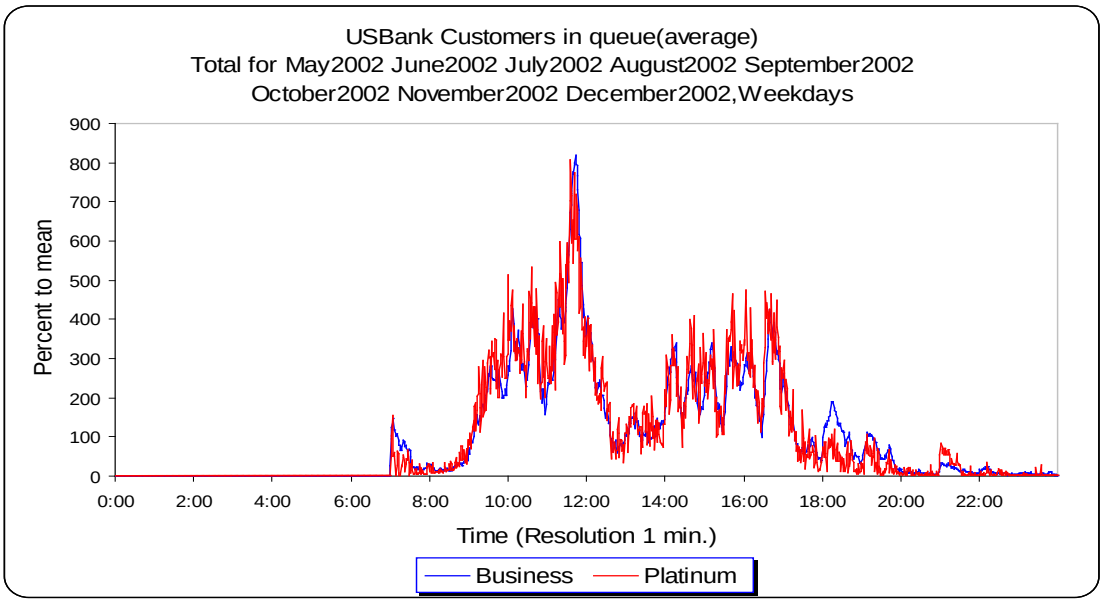
<u>Example 3.3</u>: Queue length & state-space collapse

Via **SEEStat** return to the **"Statistical Models (Summaries)"** window, click **"New Model"**. Click the **"Time Series"** button and select **"Intraday"**.
In the **"Variables"** tab select **"Customers in queue (average)"**.
In the **"Options"** tab select smoothing **"None"** and chart type **"Polygon".**
In the **"X Properties"** tab select resolution **1 minute (01:00).**
In the **"Select Categories"** tab select (with **Ctrl** and click) **Business** and **Platinum**.
Click **"Dates->"**, select **"Dates totals only",** select the 8 months from **May 2002** to **December 2002** (with **Shift**) and click) and select **Weekdays** in the **"Days"** tab.
Click **"OK"**.



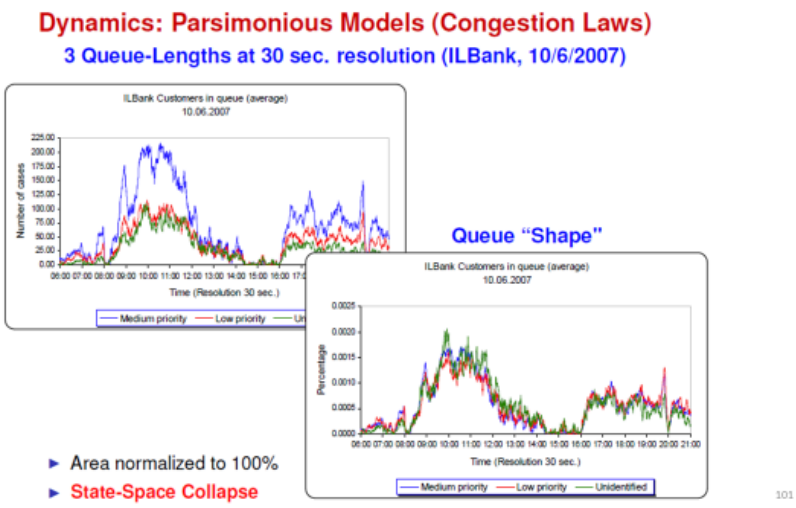
Platinum is a small-scale service. You will now normalize the chart in order to identify patterns.

Click **"Output"** on the main menu and then **"Modify Tables and Charts"**.
Open the **"Options"** tab and select **Percent to mean**. Click **"OK"**.



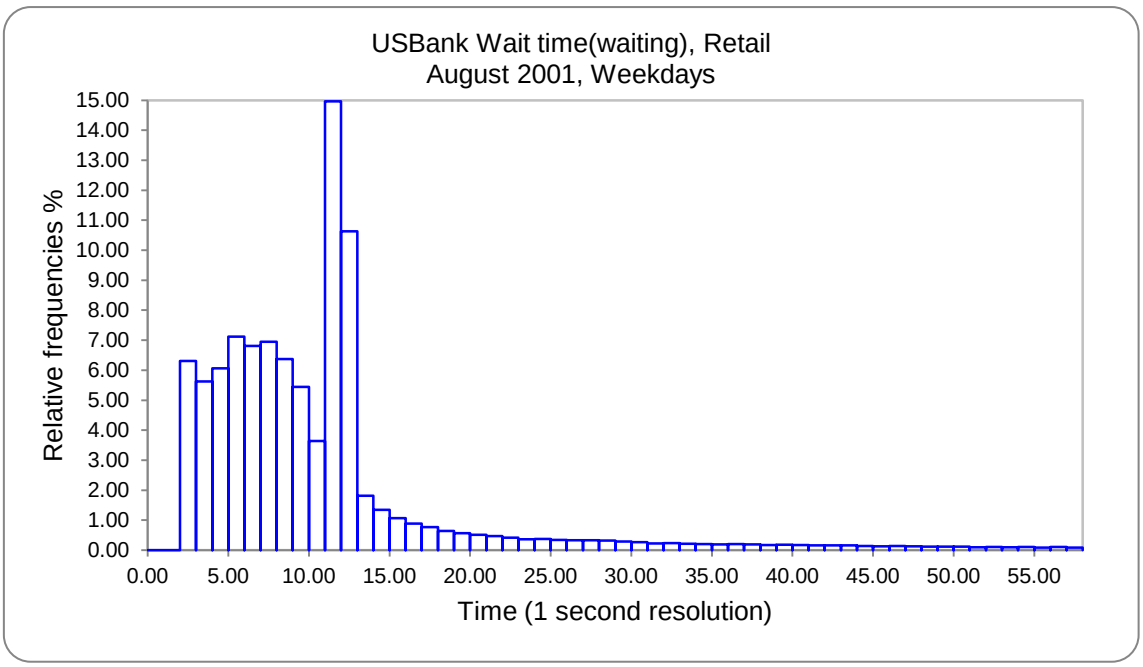
Note now the highly-overlapping patterns of the queue lengths of the two customer types. This phenomenon is predicted (not always though) by asymptotic analysis of queues in heavy traffic, where it is referred to as <u>State-Space-Collapse</u>. Here 2-dimensional stochastic-variability collapsed into a single dimension.

This state-space-collapse phenomena is in fact stronger: one can often prove that it holds at the level of individual sample-paths, as opposed to merely for averages as above. More precisely, there would be a single stochastic process that captures "shape" (the above chart represents its average). Then the random paths of the queue-process of a given type (e.g. Business) can be reconstructed via multiplying the shape-process by a <u>non-random</u> constant, with each type having its own constant (e.g. the constant of Platinum would be larger than that of Business). Here is an example from another bank, depicting 3 queues that collapse into 1:



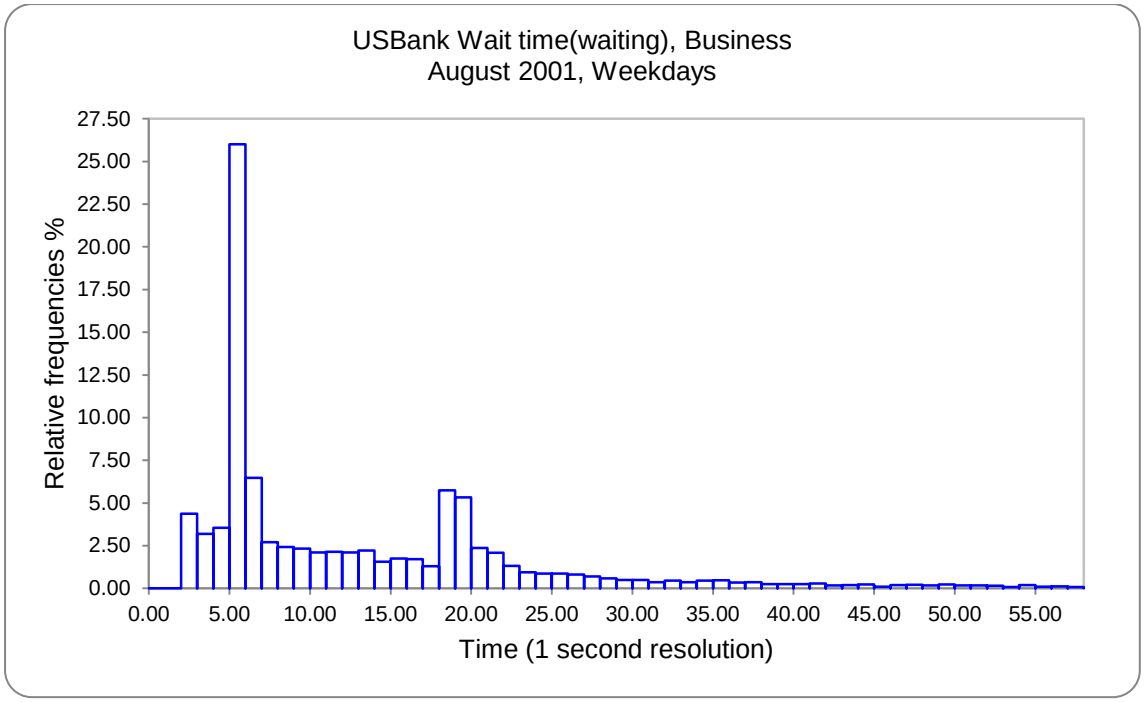
<u>Example 3.4</u>: Protocol Mining - Understanding Network Routing via SEEStat

Via **SEEStat** return to the **"Statistical Models (Summaries)"** window,
click **"Distributions"**. Click **"Estimates**.
In the **"Variables"** tab select **"Wait time (waiting)"**.
In the **"Options"** tab chart type **"Histogram".**
In the **"Select Categories"** tab select **Retail**.
In the **"X Properties"** tab select resolution **1 second (00:01).**
Click **"Dates->"** select **August 2001**and select **Weekdays** in the **"Days"** tab.
Click **"OK"**.



Via **SEEStat** return to the **"Statistical Models (Summaries)"** window, click **"<-Tables".**

In the **"Select Categories"** tab select **Business**. Click **"OK"**.



The last 2 charts present distributions of waiting-time for Retail and Business calls, during August 2001. What is the cause of the observed peaks?

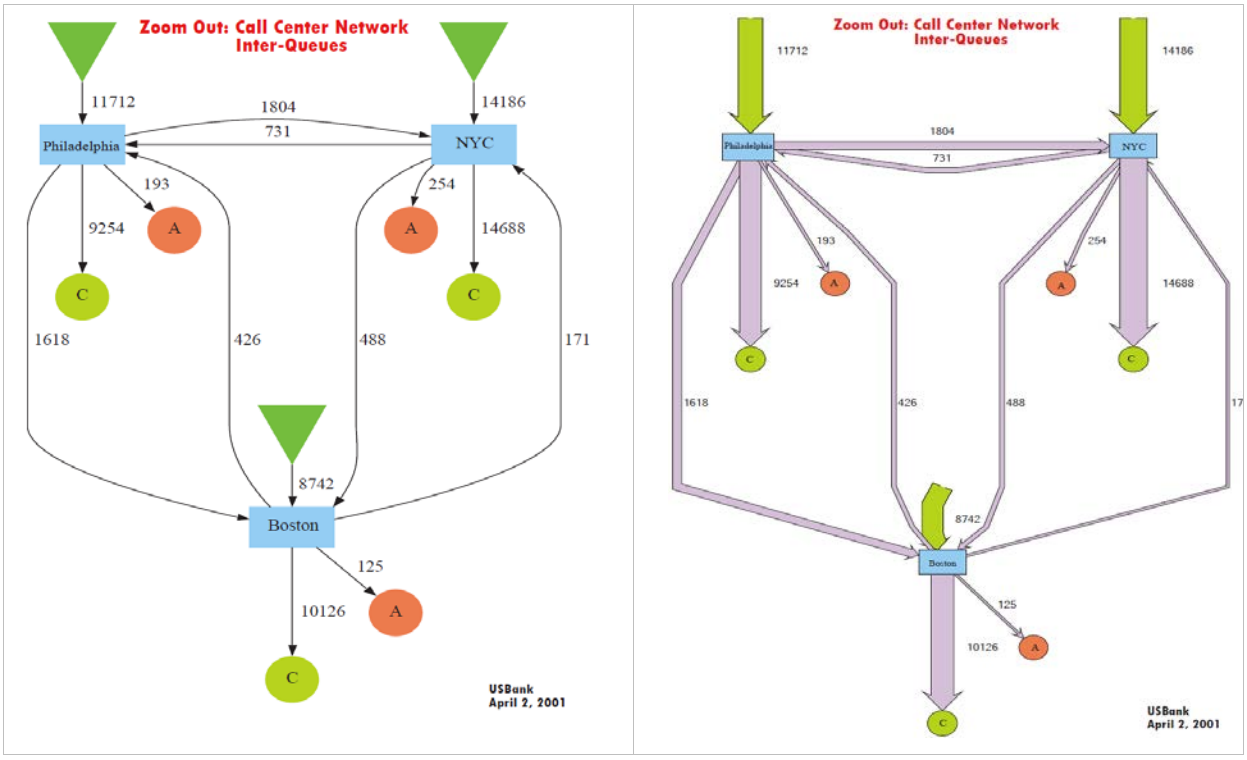
For Retail customers, the peak is a little above 10 sec: the reason is a ***routing protocol*** *that USBank applies to Retail customers. To understand it, one must become aware of the fact that the call center of USBank in facts consists of 3 geographically-distributed yet logically-centralized call centers: in NYC, Boston and Philadelphia. (See the two SEEGraphs below, which provide useful examples of "structure-mining".) Now suppose that a Retail customer in NYC calls; then this customer joins the NYC queue; if that customer is not served within 10 seconds, then s/he is "transferred" to (what is called) an* ***inter-queue****, which is a virtual queue that is being served by all 3 call centers; then that customer will be served by the next feasible agent at any of the 3 call centers.*

Why do that? This is effectively dynamic load-balancing, or pooling, which seeks to take (partial) advantage of the operational benefits associated with operating a larger-system (economies-of-scale). This also explains why not do it, say, after 5 minutes.

But why not do it immediately "after 0 seconds"? Since the communication system at the time would have collapsed due to additional information transfer (e.g. all 3 call centers must remain continuously updated about the status of say NYC customers).

For Business calls, the peak at 5 seconds is for the same reason. Since Business customers better enjoy higher levels of service than Retail, the threshold was reduced to 5 seconds.

The second peak, at 18 seconds, is unclear – it can possibly be the outcome of a priority-upgrade, locally (in NYC, for the above example) or globally (in the Inter-queue).



Few words about SEEGraphs, or Structure-Mining

The above two SEEGraphs are the outcome of a process that I refer to as "Structure-Mining": here, the outcome of this process are graphs that depict (logical) flow of customers among the three call centers (NYC, Philadelphia, Boston).

Both graphs tell the same story. Consider, for example, the Philadelphia call center during April 2: 11712 called it, 9254 got served, 193 abandoned, 1804 transferred to the inter-queue and got served in NYC and 1618 transferred and served in Boston; Philadelphia also served 426 from Boston and 731 from NYC.

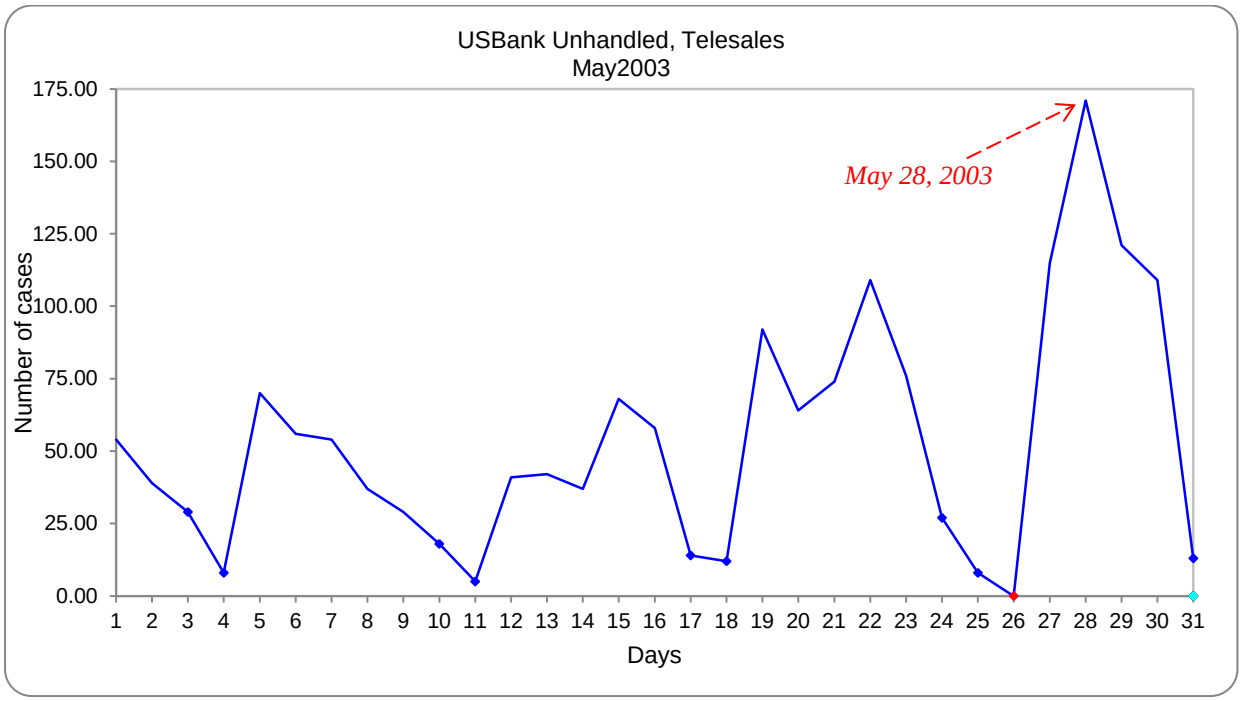
Flow volumes are easier to compare qualitatively via the right SEEGraph – the higher the volume the thicker the arc-width.

SEEGraphs (and SEEnimations, their dynamic counterparts) are described briefly in Appendix A. (They are the focus of a second tutorial, which is still in the making.) As will be clear when reading Appendix A, Structure-Mining is a prerequisite for the creation of essentially every SEEGraph or SEEnimation.

<u>Example 3.5</u>: Protocol Mining - Discovering and Understanding an Operational Flaw via SEEStat

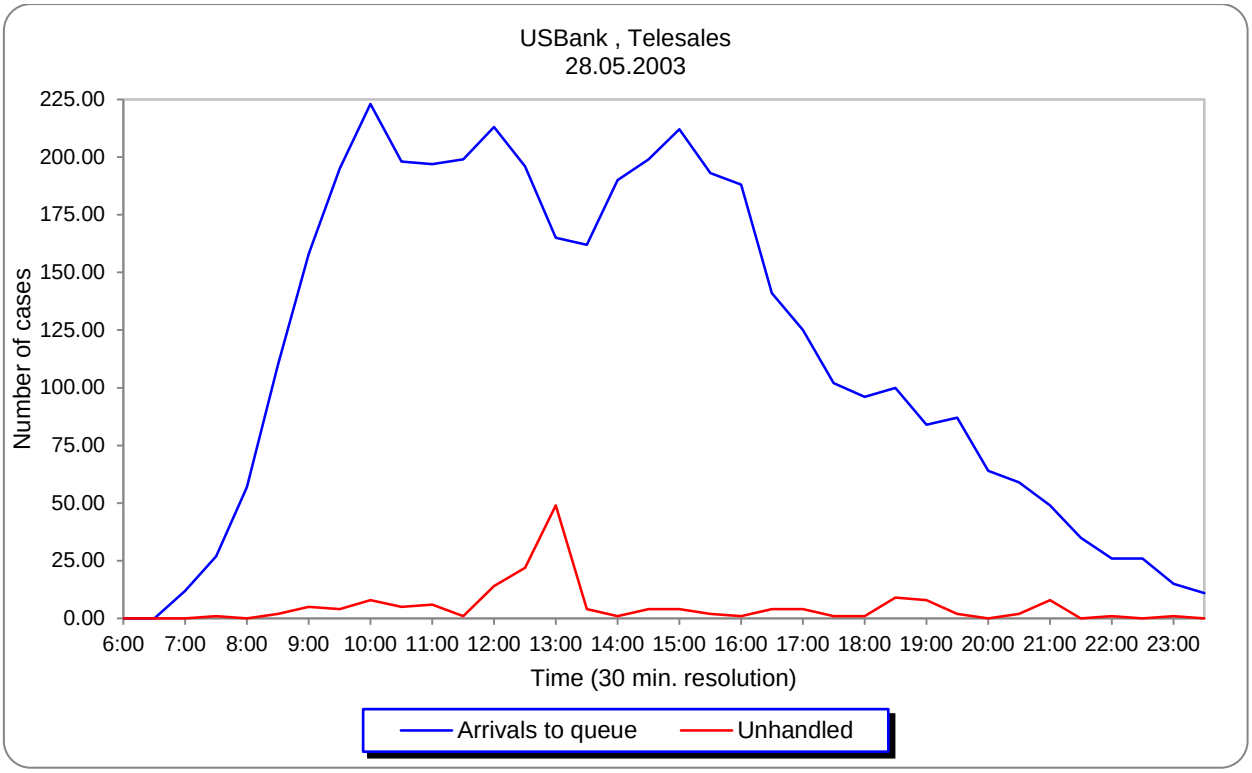
Analyzing peaks of Telesales unhandled, May 2003

Click **"Main"**, and select **"Statistical Models (Summaries)".** Select **"New Model"**, **"Time Series"**, and then **"Daily totals"**. In the **"Variables"** tab select **"Unhandled"**. In the **"Select Categories"** tab select **Telesales**.
Click **"Dates->"** select **Days for one month,** and **May 2003.** Click **"OK"**.



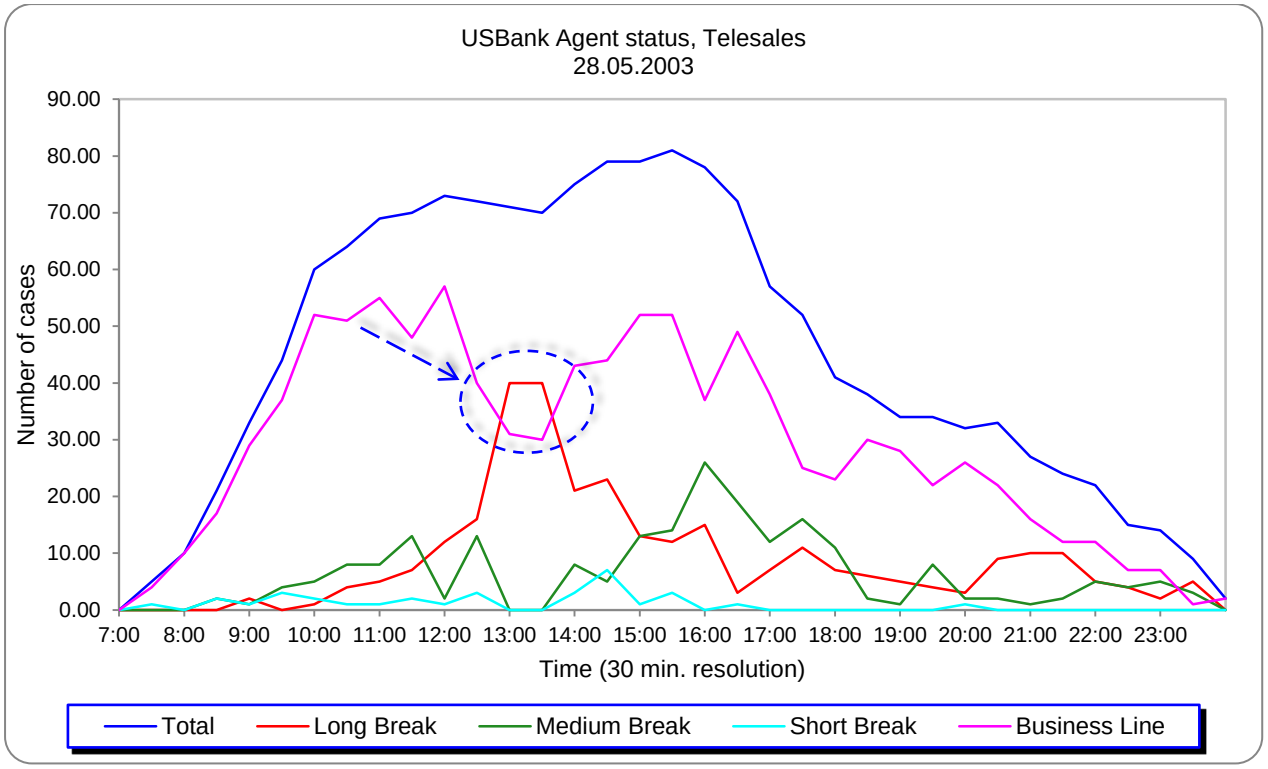
A monthly picture identifies a peak in the unhandled Telesales calls, on Wednesday, May, 28, 2003. We shall now focus on that day.

Return to the SEEStat, click **Windows** and select **"Statistical Models (Summaries)"** window. Click **"New Model".** Select **"Time Series"**-> **"Intraday".**
In the **"Variables"** tab select (using **Ctrl**) both **"Arrivals to queue"** and **"Unhandled".**
In the **"Select Categories"** tab select **Telesales**.
Open **"Properties"**, select 30 minutes resolution (30:00), and change **Low Limit** to **06:00.**
Click **"Dates->".** Select **"Individual Days"** and **May 2003**. Click **Days** tab and select **28 May**.
Click **"OK"**.



We observe a peak of unhandled calls in 13:00, simultaneously with a significant decrease of the arrival rate.

Return to SEEStat, click **Windows** and select **"Statistical Models (Summaries)"** window. Click **"<- Tables".** In the **"Variables"** tab select **"Agent status"**.
In the **"Select Categories"** tab select **Telesales**, and all subcategories *except NonBusiness Line.*
Open **"Properties"**, select 30 minutes resolution (30:00), and change **Low Limit** to **07:00.**
Click **"OK"**.

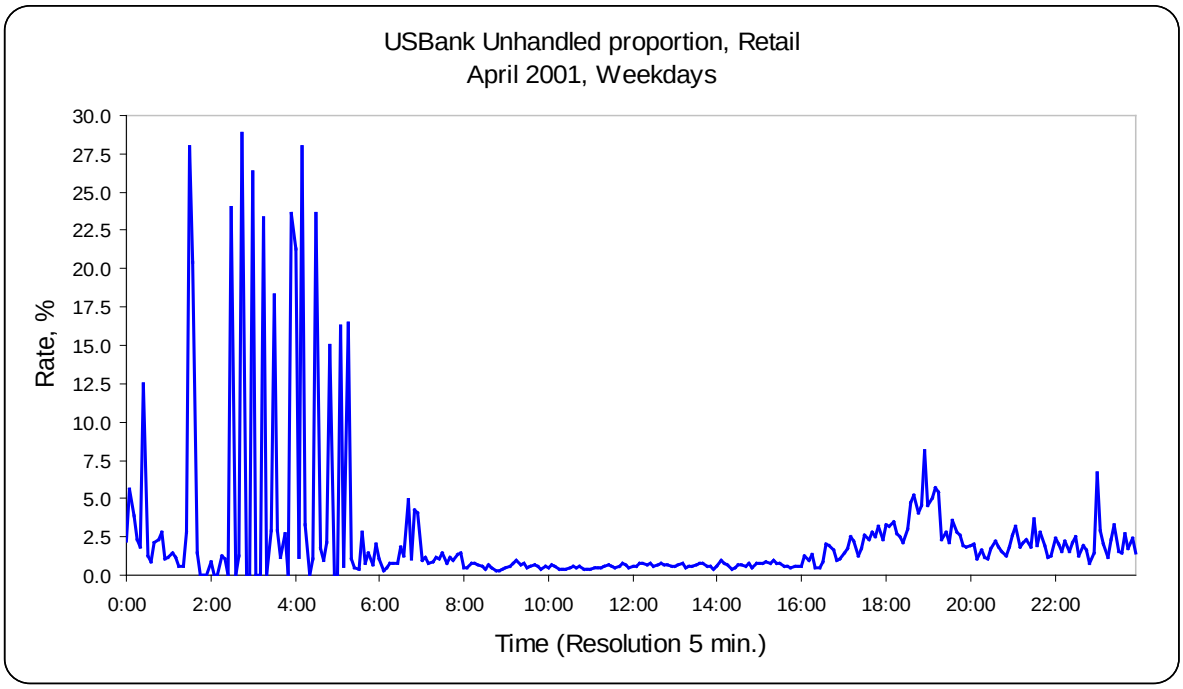


Using Agent Status of Telesales, we observe that there was a significant decrease in the number of agents serving incoming calls, between 12:00 and 13:00 (decrease from 57 to 31), simultaneously with a sharp increase in the number of agents who were on a long-break.

Final diagnosis: conceivably, the deterioration in service-level (increase in unhandled) stems from mismanaging staffing levels during or around the lunch break. Indeed, we saw a decrease in the arrival rate at and right after 13:00, which does warrant a decrease in staffing levels. However, the actual decrease (e.g. agents leaving on long breaks) led to understaffing and hence over-congestion.

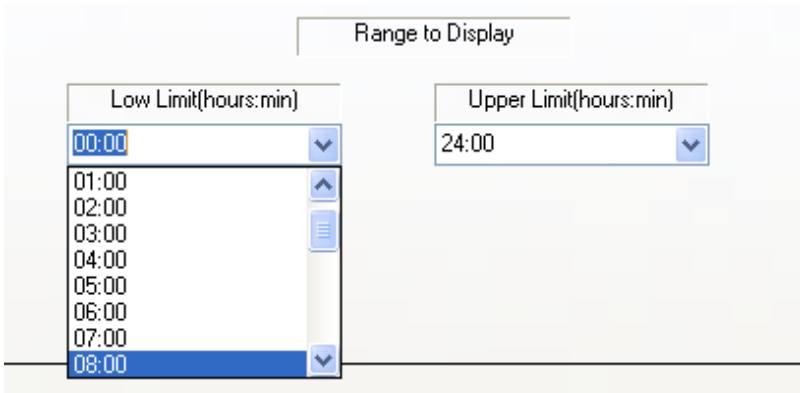
<u>Example 3.6</u>: Change-of-Shifts phenomena (or, staffing levels vs. Offered-Load)

Click **"Main"**, and select **"Statistical Models (Summaries)".** Select **"Time Series"**, then **"Intraday "**.
In the **"Variables"** tab select **"Unhandled proportion"**.
In the **"X Properties"** tab select resolution **5 minutes.** In **"Select Categories"** tab select **"Retail".** Click **"Dates->"**, select **"April 2001"** and **"Aggregated Days",** and select **Weekdays**. Click **"OK"**.

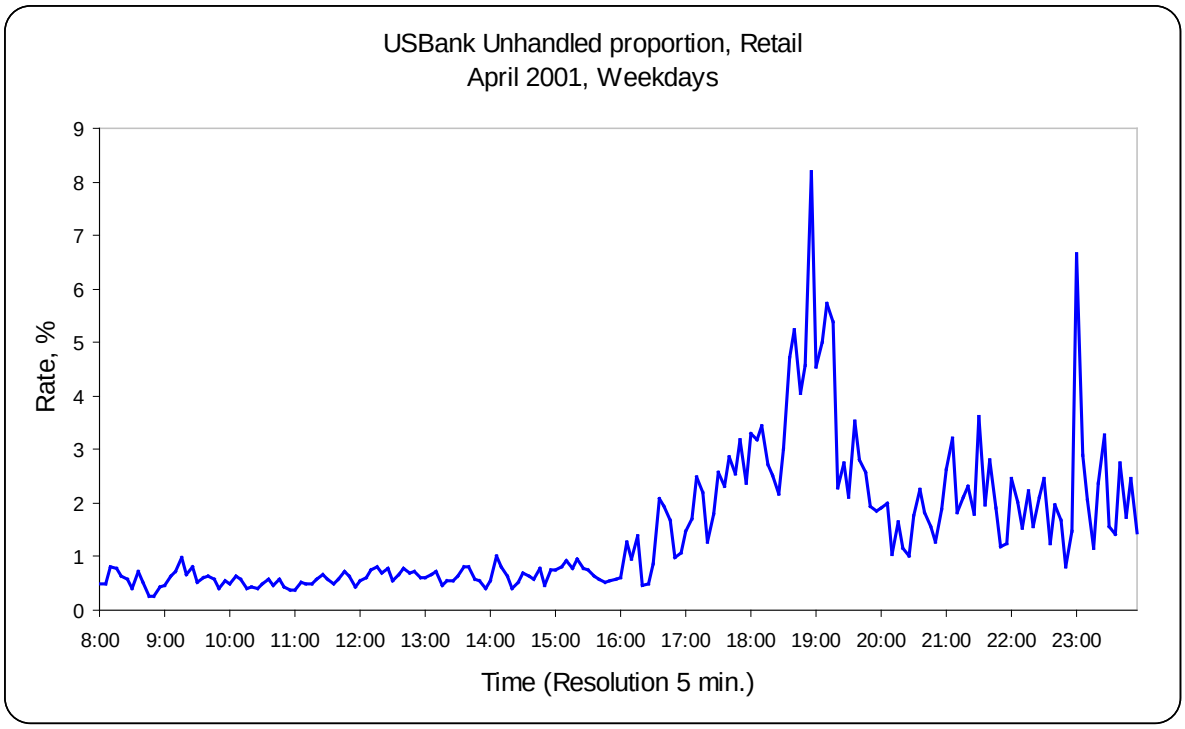


We observe a lot of noise before 8:00 a.m. There are only a few agents working then, and few customers are calling. We now cut this noisy (possibly irrelevant) part of the chart, until 8:00 a.m.

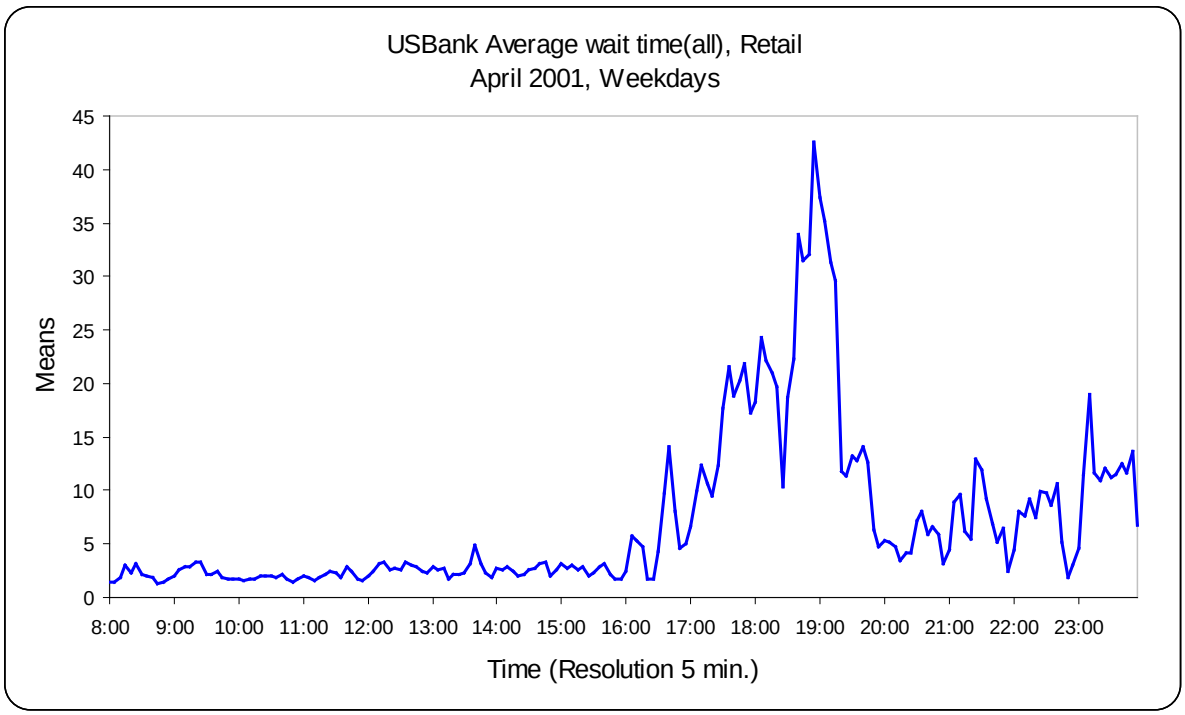
Click **"Output"** on the main menu and then **"Modify Tables and Charts"**.
Open the **"Properties"** tab and change low limit to **08:00.**



Click **"OK"**.

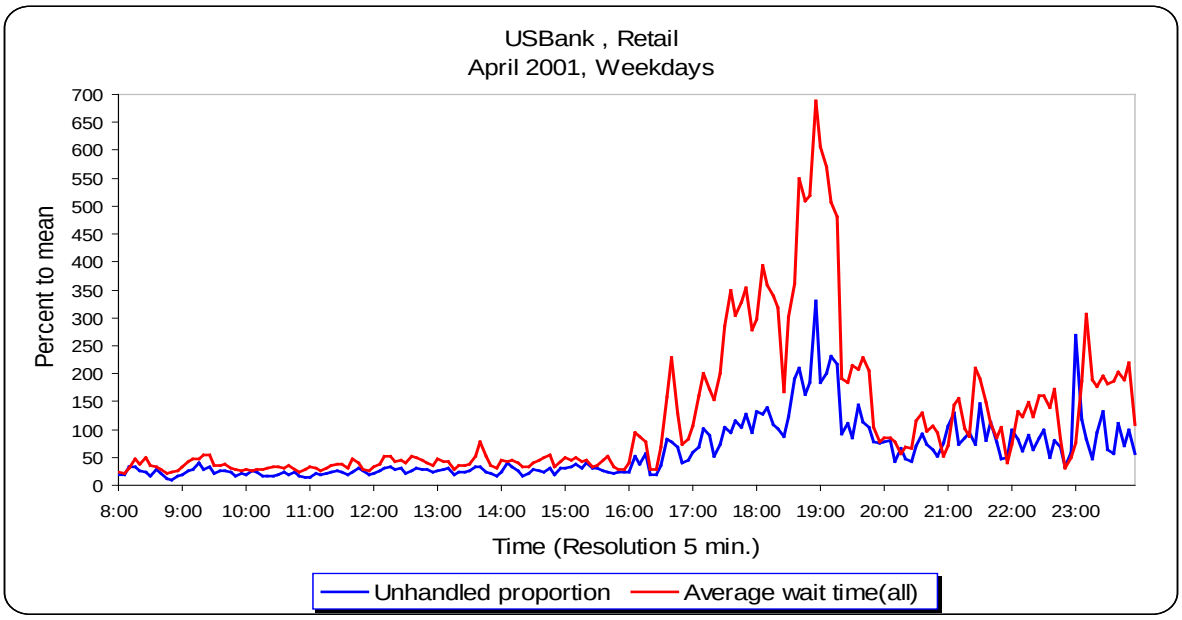


Via **SEESTAT** return to the **"Statistical Models (Summaries)"** window.
Click **"<-Tables"**. In the **"Variables"** tab select **"Average wait time (all)".** Click **"OK"**



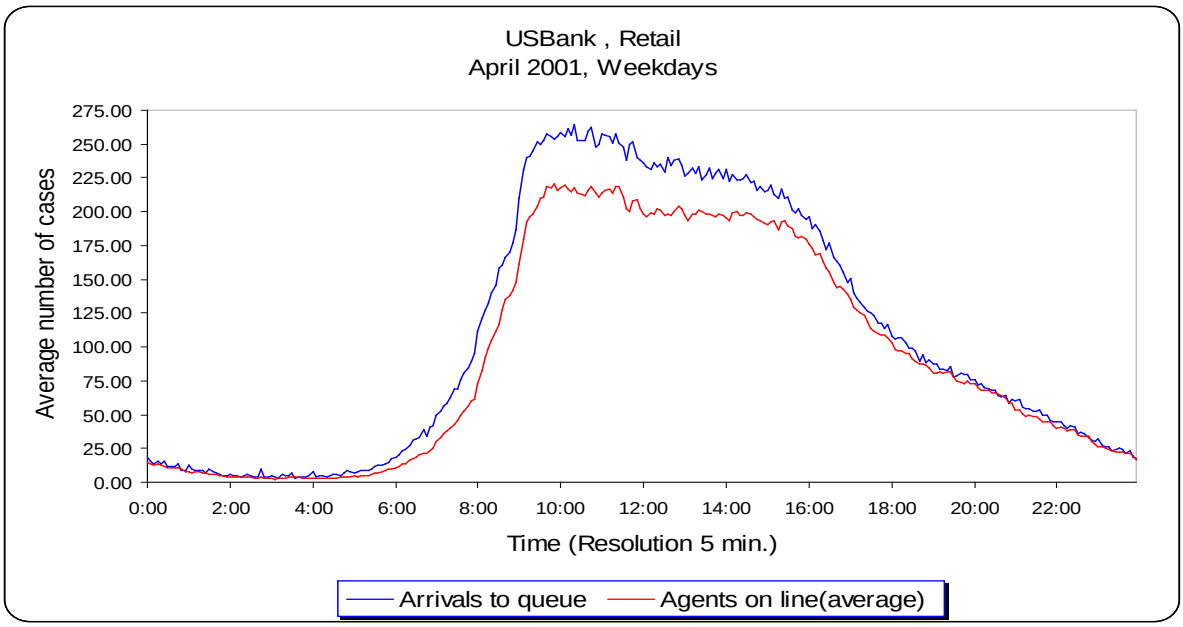
Note that the patterns for the two variables ("Unhandled proportion" and "Average wait time (all)" are rather similar. We now compare them more closely.

Via **SEESTAT** return to the **"Statistical Models (Summaries)"** window. On tab **"Variables"** select **"Unhandled proportion"** and **"Average wait time (all)"**. Click **OK.**



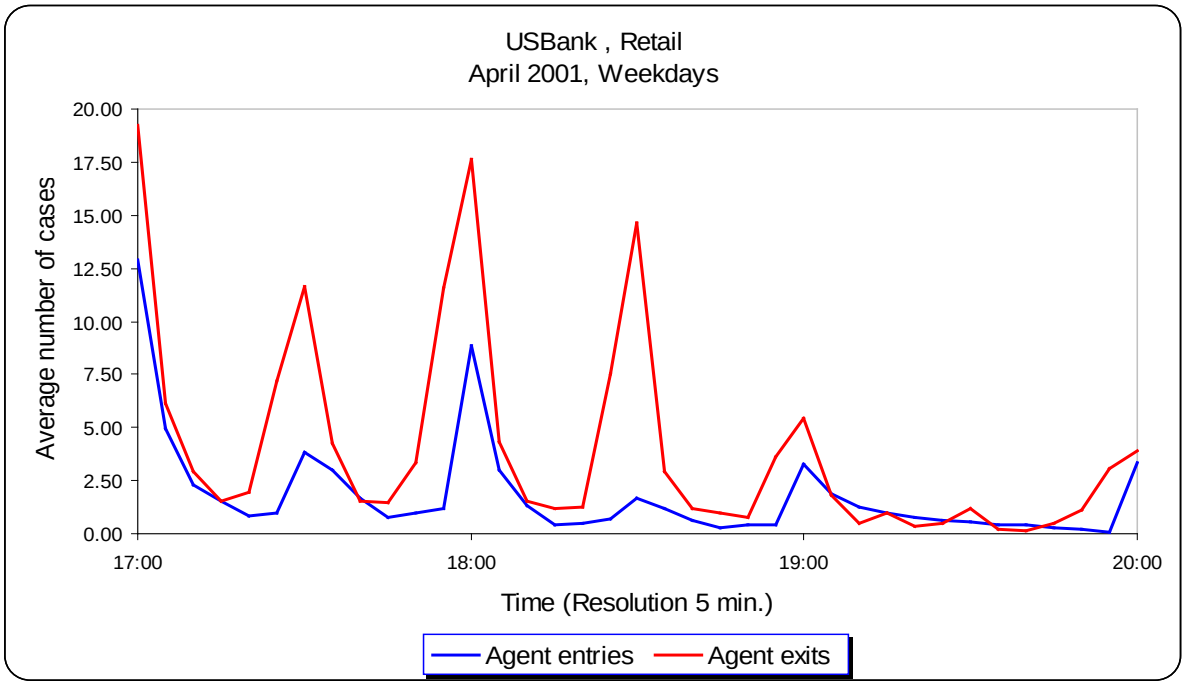
Note an increase in "unhandled proportion" and "average wait time" from 17:00 to 20:00 – this is a time period of shift-change, or shift-overlap. Indeed, we shall verify, momentarily, that during this period, many agents were leaving/completing their shifts. The number of arrivals is also going down, but the schedule of agent exits does seem to be well-synchronized with arrivals (seems like agents here are leaving prematurely).

With SEEStat, we created for you the following chart: customer arrivals and agents online; but this does not provide a conclusive evidence for what seems to be happening. To get that evidence, we now dig deeper.



The management of shift-change is a prevalent chronic problem for call centers. We identify such problems via the SEEStat functions that display entries and exits of agents:

Via **SEESTAT** return to the **"Statistical Models (Summaries)"** window. In the **"Variables"** tab select **"Agent entries"** and **"Agent exits"**. In the **"Select Categories"** tab select **"Retail".** Open the **"Properties"** tab and change the low limit to **17:00** and upper limit to **20:00.** Click **OK.**



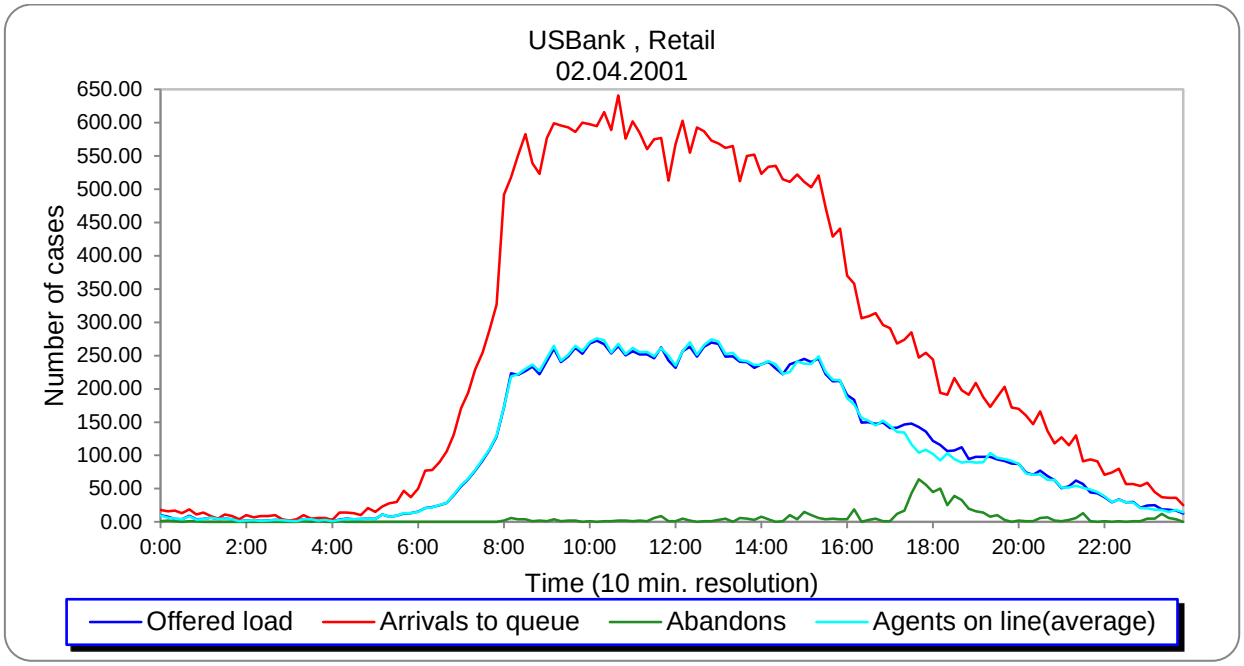
Note that times of entries immediately follow times of exists (some overlap would have been desirable). In addition, more are leaving (in red) than are joining (blue) which, as noted, was not matched well with the decline rate of customer calls.

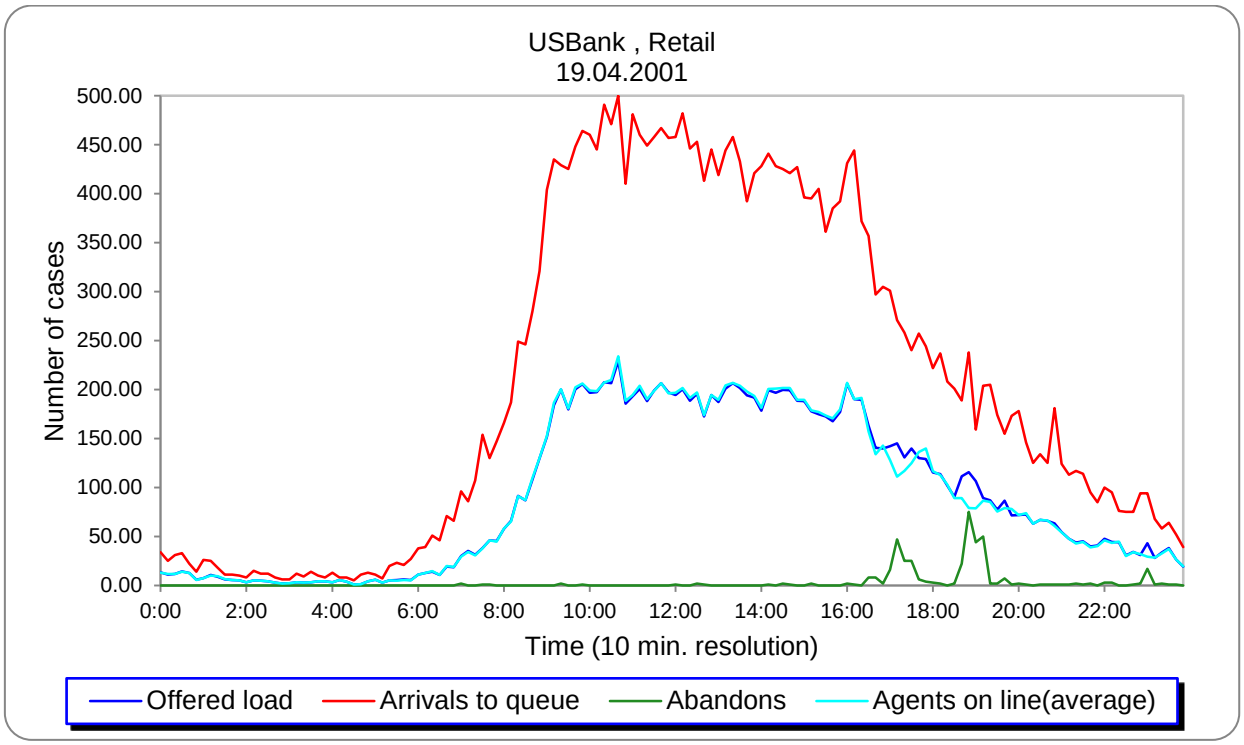
Mismatch between staffing levels and the offered-load: *In order to truly understand what is happening in the above, one must refine the analysis beyond monthly averages. To this end, we focus on 2 relatively-busy days (02.04.2001 and 19.04.2001), and plot for them intraday arrivals, abandonment, agents-on-line and what we call the* ***offered-load*** [3]*.*
As clear from the 2 next charts, the number of abandonment significantly and visibly increases when there is a noticeable gap between the number of agents-on-line and the offered-load (the latter being above the former).
This happens during a single period around 18:00 for one day, and over 2 separate periods around the same time for the second day. (Recall that change-of-shift time is around 18:00).

Finally, note the high data-resolution and the sophistication required to clearly understand the cause for service-level deterioration around 18:00; one should also note that the offered-load not only helps explain what is going on, but it also quantifies how many agents are actually required so as to avoid such deterioration (not that many more

[3] *In Appendix G, we shall briefly discuss, and provide further readings for, the key concept of* ***offered-load****. For now, conceptualize it as the amount of work, in units of number-of-agents, that is imposed on (required from) the system; it is calculated by combining arrival rates and service durations (e.g. the same offered-load can arise from many arrivals that require short services or from few arrivals that require each a long service). One reason for the importance of the offered-load is that appropriate staffing levels reside around it (much above (below) gives rise to high (low) service level).*

in the 2 days below, as can be read-off the Excel tables associated with the 2 graphs below).



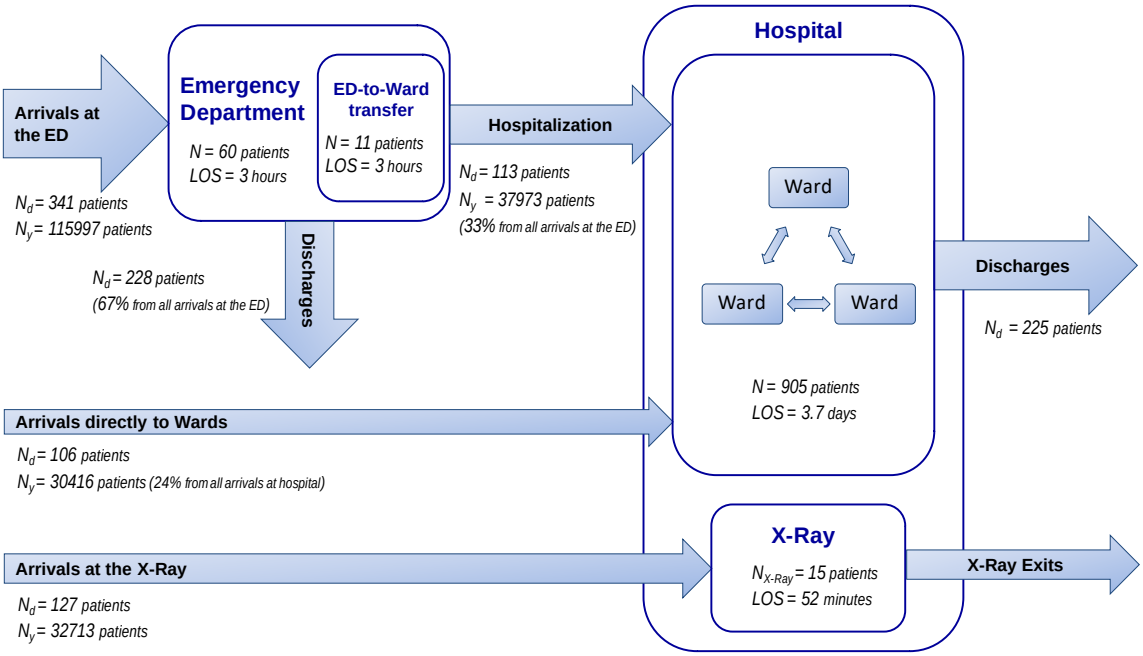


Exit SEEStat, either via the "**X**" on the top-right corner, or by clicking "**Close SEEStat**" in the Main menu. (Don't exit the Terminal.)

We shall now continue with the EDA (Exploratory Data Analysis) of a second dataset – taken from an **Israeli hospital** that has been a SEELab data-partner. This will allow you to gain further experience with SEEStat and enjoy (so is our hope) more examples.
An additional important goal is to amplify the fact that our EDA is relevant beyond one specific service operation or one specific service industry – it is in fact a prerequisite for the engineering of any service (hence Service Engineering) – be it a call center, or an emergency department or ….

HomeHospital Data

***Background**: The data we rely on was gathered at a data-partner hospital, which is a large tertiary hospital in Northern Israel. This hospital consists of about 1000 beds and 45 medical units. The data covers detailed information (mostly operational) on patient-flow throughout the hospital, over a period of several years (January 2004–October 2007). In particular, the data allows one to follow time-stamped paths of individual patients throughout their hospital journey, including admission, discharge, and transfers between hospital units. The data is inter-departmental in the sense that it does not acknowledge resolutions within the ED or within wards.*



N_d - average number of patients that arrived per weekday, for period January 1, 2004–October, 31, 2007

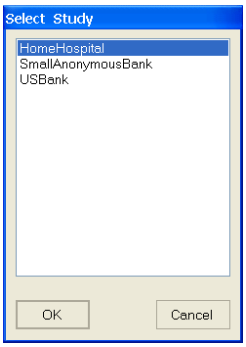
N_y - average number of patients that arrived per year, for years 2004, 2005, 2006, all days (for year 2007 data not fully completed; missing two months—November and December).

N - average number of patients in ED/ED-to-Ward transfer/Wards, recorded at 12:00 per weekday, for period January 1, 2004–October, 31, 2007

$N_{X\text{-}Ray}$ - average number of patients in X-Ray at 10:00 per weekday, for period January 1, 2004–October, 31, 2007

LOS - length of stay in ED/ED-to-Ward transfer/Wards/X-Ray per weekday, for period January 1, 2004–October, 31, 2007

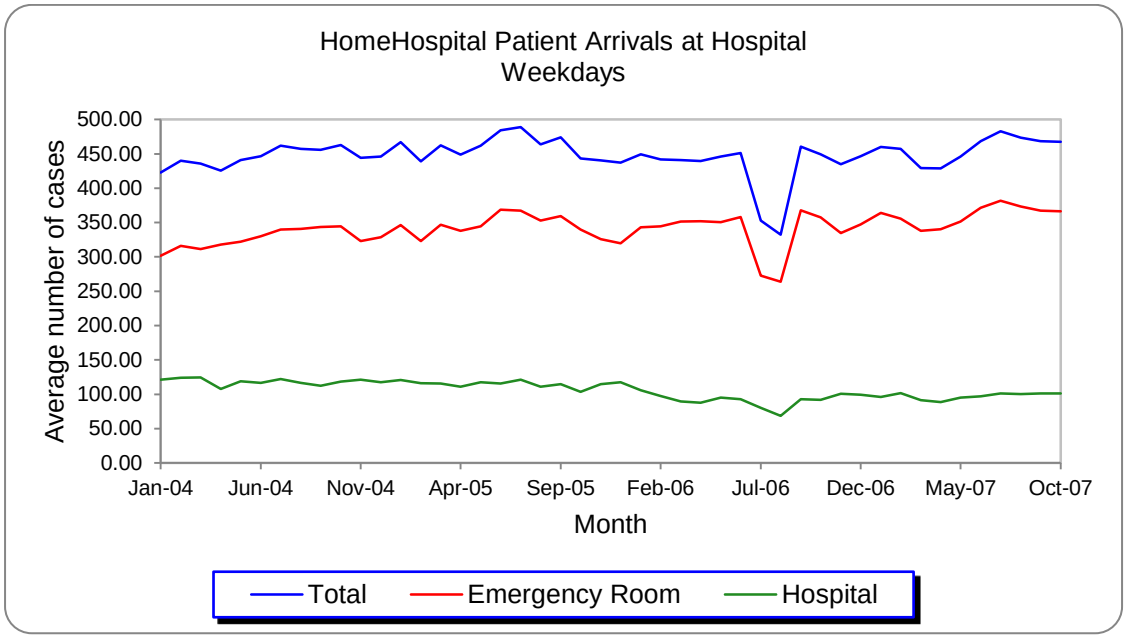
Reopen SEEStat 3.0 and select the **HomeHospital** study.



Part 4: Hospital

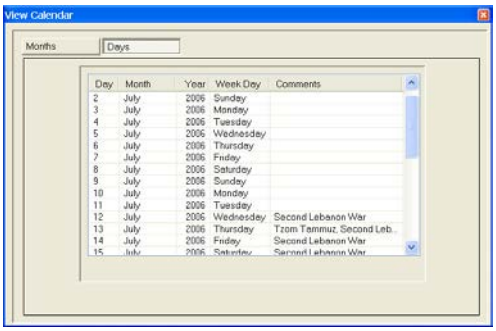
Example 4.1: *Arrivals - Average per one weekday over entire month*

Click **"Main"** and **"Statistical Models (Summaries)".** Select **"Time Series"**, then **"Daily totals"**.
From the variables list select **"Patient Arrivals at Hospital".**
In the **"Select Categories"** tab, select all categories.
Click the **"Dates->"** button. Click the **"Select all"** button, open tab **"Days"** and select **"Weekdays"**.
Click **"OK"**.



*Note the drop in the number of arrivals during July and August 2006. The reason for this phenomenon is the "second Lebanon war": it took place during that period and affected mostly the northern part of Israel in which the hospital is located. This could be verified by clicking "**Calendar**", in which special days (such as holidays) and special events (wars☹) are noted:*

Click **View**-> **Calendar**. Mark **"Individual days"** and select **July 2006.** Open tab **"Days"**.

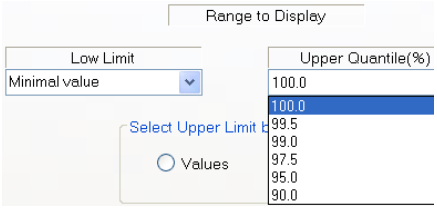


Click **"Months"** tab and select **August 2006.** Open tab **"Days"**.

Part 5: Emergency Department

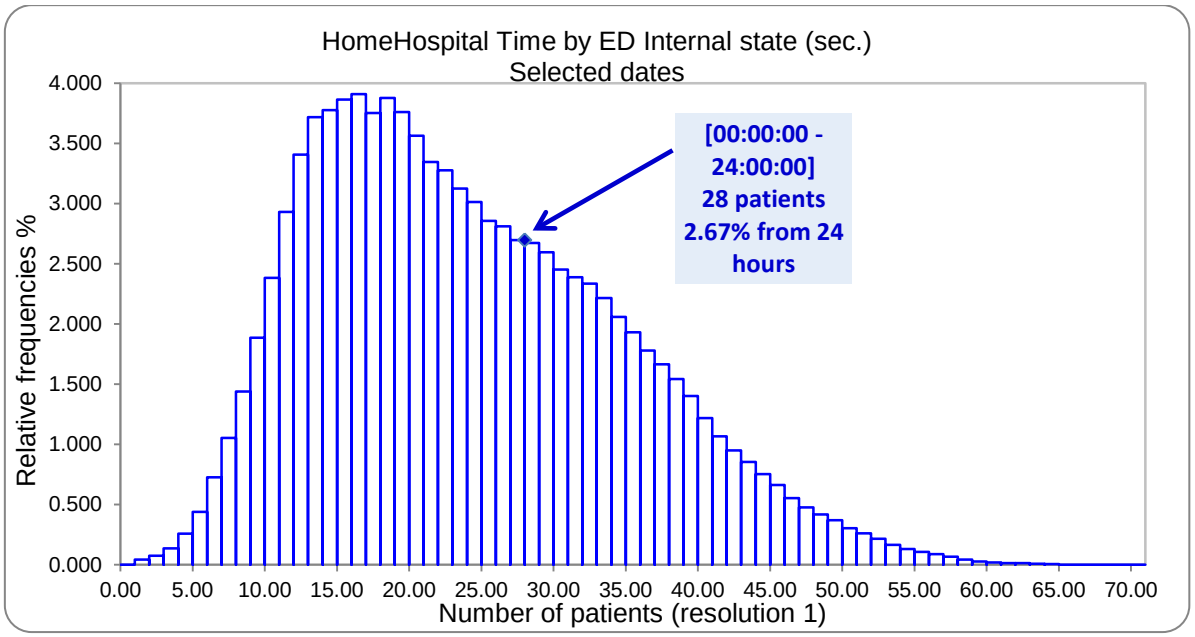
***Example 5.1*: Distribution of ED Occupancy, overall. (Time by ED Internal state (sec.), or equivalently ED Census Distribution during all 24 hours of the day.)**

Return to the **"Statistical Models (Summaries)"** window.
Click the **"New Model"** button. Select **"Distributions"**, then **"Estimates"**.
In the **"Variable"** tab, select **"Time by ED Internal state (sec.)"**. (This variable, as seen momentarily, has a rather complex meaning and a very interesting behavior.)
In the **"Select Categories"** tab, select **"Total"**.
In the **"X Properties"** tab change **upper quantile** limit to **100.**



Click the **"Dates->"** button. Select **"Dates totals only"** and all months from **January 2004 to October 2007,** open the **"Days"** tab and select **"All days"**.
Click **"OK"**.

Remark on DB Management in SEEStat: SEEStat's fast response time (measured in seconds) enables online EDA. This is notable since, for example, the above query searches through data that occupies over 3.5GB disk space. (The previous project of USBank occupies close to 7GB.) There is more, however: the present tutorial is querying only SEEStat's *data summary tables* (all in binary format), but these were created from complete log-files that had been imported from our data-partners (and transformed into often very large MS Access data bases, specifically of size 50GB for USBank and even 180GB in other cases. Notably, the original data for HomeHospital occupied only 343 MB – it grew to 3.5BG with the many variables, and hence data-tables, that were created to support EDA. A glimpse to these variables will appear in the sequel.



*For example: 28 patients were in Internal ED during 38 minutes and 30 seconds (2.674% from 24 hours: 0.02674*86400 sec=2310 sec) between 00:00 and 24:00.*

Remark: The fraction 2.674% is in fact calculated from the full sample which here, as will be now explained, consists of about 120 million seconds.

The following is EXCEL's **Table**, *created jointly with the above* **Chart**:

Statistics	
	Time by ED Internal state (sec.)
N	120960000
N(average per day)	86400
Mean	23.63
Standard Deviation	10.7
Variance	114.4
Median	22
Minimum	0
Maximum	70

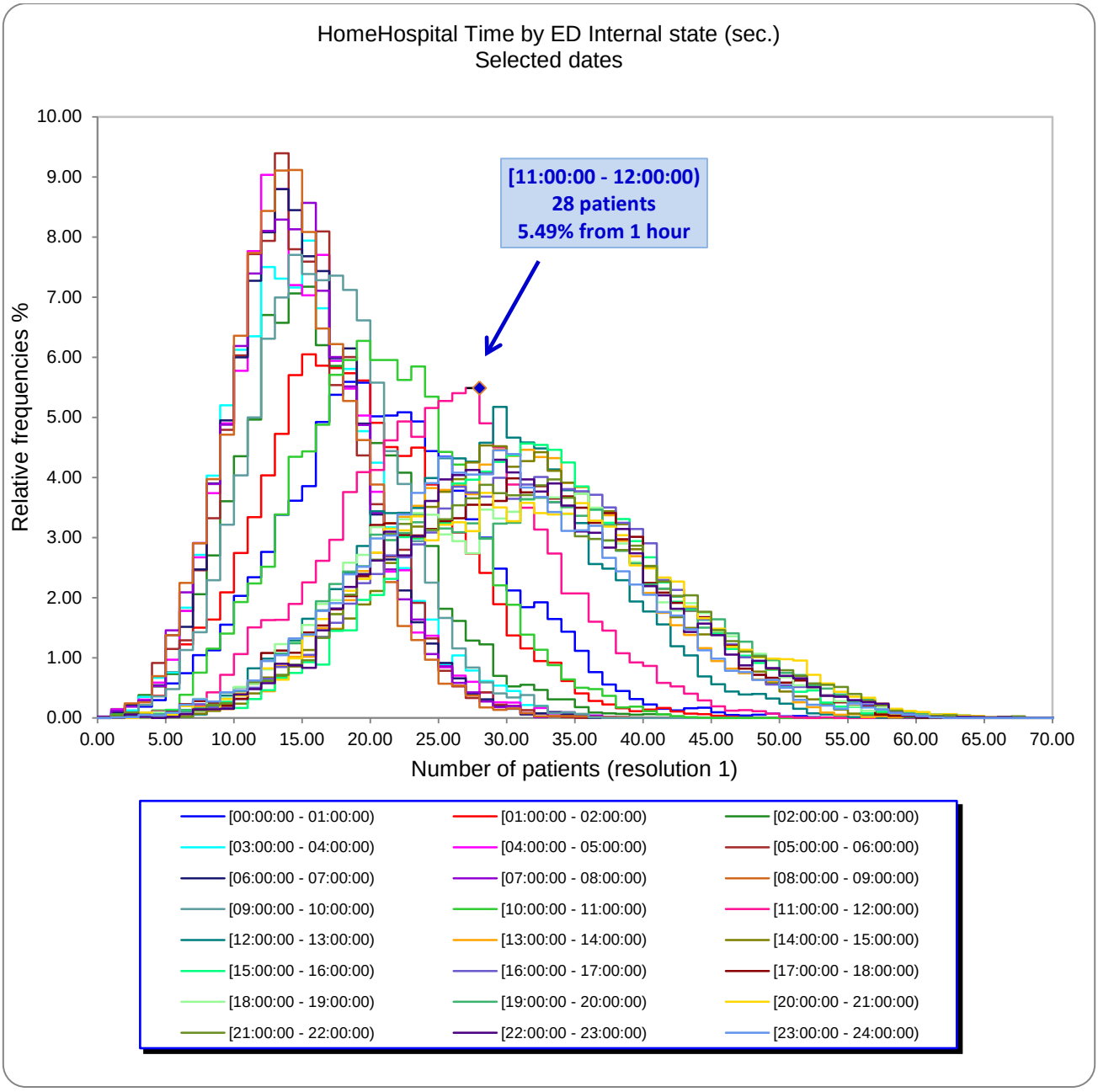
*Explanation of sample size in the above table: N=120960000 (1400 days*86400 seconds), namely around 120 million seconds, and N (average per day) = 86400 seconds (60 minutes*60 seconds*24 hours). These specific numbers arise from the way SEEStat calculates instants counts (for each moment of time (1 seconds): calculate the system-state (in this case number of patients in ED); and then calculate frequencies (here it is the time) of instants counts). For more detail on instant counts, see pages 49-50 in* Reproducing EDA via SEEStat (Link).

***Note**: Denote by L the ED occupancy, and view L as a random variable. Then the above graph is simply the empirical histogram/distribution of L, estimated from all of our hospital data. In particular, by the latter Statistics table (Table 1), an estimate for the mean of D is 23.63 beds and for the standard deviation 10.7.*

We observe that the occupancy distribution has a somewhat unusual shape; it is skewed to the left, has a light right tail, has one clear local peak (around 15) and what seems to be one inflection point (around 28). We shall now further investigate this distribution via the following Example 5.2.

***Example 5.2*: Distribution of ED Occupancy, separately for each of the 24 hours in a day. (Time by ED Internal state (sec.), or equivalently ED census distribution during each of the 24 hours of the day.)**

Return to the **"Statistical Models (Summaries)"** window.
Click the **"<-Tables"** button.
In the **"Select Categories"** tab, select (with shift key) all categories except **"Total"**.
Click **"OK"**.



*For example: 28 patients were in the Internal ED during 3 minutes and 17 seconds (5.49% from 1 hour: 0.0549*3600 sec=197 sec) between 11:00 and 12:00.*

Here is part of the **Table** *that was created jointly with the above* **Chart***:*

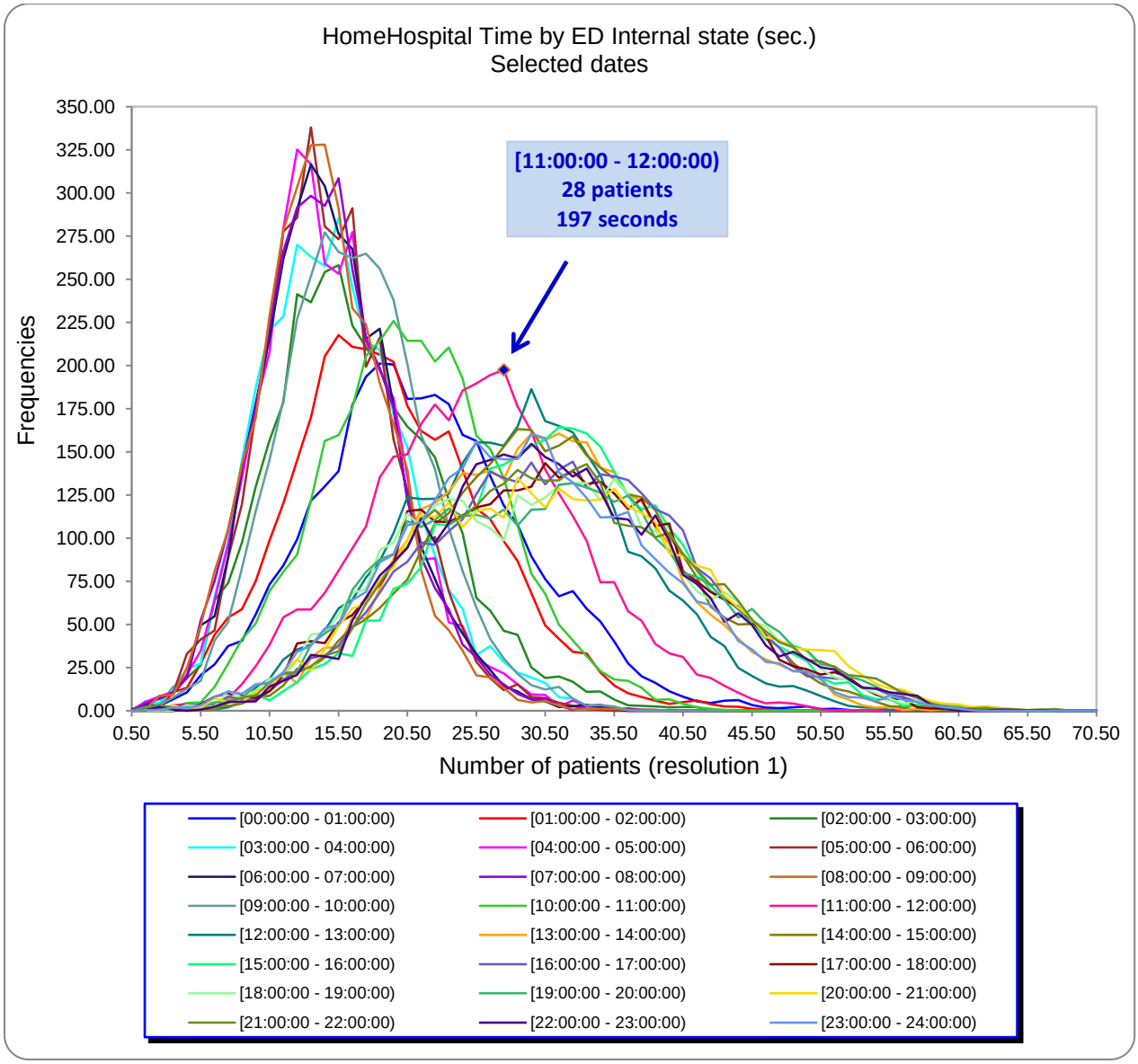
Statistics

	N	N(average per day)	Mean	Standard Deviation	Variance	Median	Minimum	Maximum
[00:00:00-01:00:00)	5040000	3600	20.93	7.6	58.09	20	0	52
[01:00:00 - 02:00:00)	5040000	3600	18.74	6.9	48.6	18	0	48
[02:00:00 - 03:00:00)	5040000	3600	16.35	6.1	37.67	16	1	42
[03:00:00 - 04:00:00)	5040000	3600	14.86	5.3	29.15	14	1	40
[04:00:00 - 05:00:00)	5040000	3600	14.43	5.0	25.26	14	1	35
[05:00:00 - 06:00:00)	5040000	3600	14.45	4.9	24.05	14	1	34
[06:00:00 - 07:00:00)	5040000	3600	14.38	4.8	23.96	14	1	36
[07:00:00 - 08:00:00)	5040000	3600	14.24	4.9	24.55	14	1	38
[08:00:00 - 09:00:00)	5040000	3600	14	4.7	22.62	14	1	39
[09:00:00 - 10:00:00)	5040000	3600	16.17	5.1	26.95	16	1	36
[10:00:00 - 11:00:00)	5040000	3600	20.36	6.4	41.23	20	1	44
[11:00:00 - 12:00:00)	5040000	3600	24.75	7.6	59.01	25	1	53
[12:00:00 - 13:00:00)	5040000	3600	28.02	8.4	71.95	28	2	67
[13:00:00 - 14:00:00)	5040000	3600	29.73	8.8	78.45	30	4	65
[14:00:00 - 15:00:00)	5040000	3600	30.59	8.9	80.57	30	4	62
[15:00:00 - 16:00:00)	5040000	3600	31.2	9.1	83.76	31	4	62
[16:00:00 - 17:00:00)	5040000	3600	31	9.6	93.63	31	4	64
[17:00:00 - 18:00:00)	5040000	3600	30.2	9.8	97.69	30	4	64
[18:00:00 - 19:00:00)	5040000	3600	30.16	10.2	104.93	30	3	64
[19:00:00 - 20:00:00)	5040000	3600	30.62	10.5	110.18	30	2	64
[20:00:00 - 21:00:00)	5040000	3600	31.14	10.5	110.41	31	2	67
[21:00:00 - 22:00:00)	5040000	3600	31.08	10.16	103.26	31	1	70
[22:00:00 - 23:00:00)	5040000	3600	30.58	9.7	94.25	30	1	66
[23:00:00 - 24:00:00)	5040000	3600	29.02	9.5	91.04	29	1	61

Click **"Output"** on the main menu and then **"Modify Tables and Charts"**.

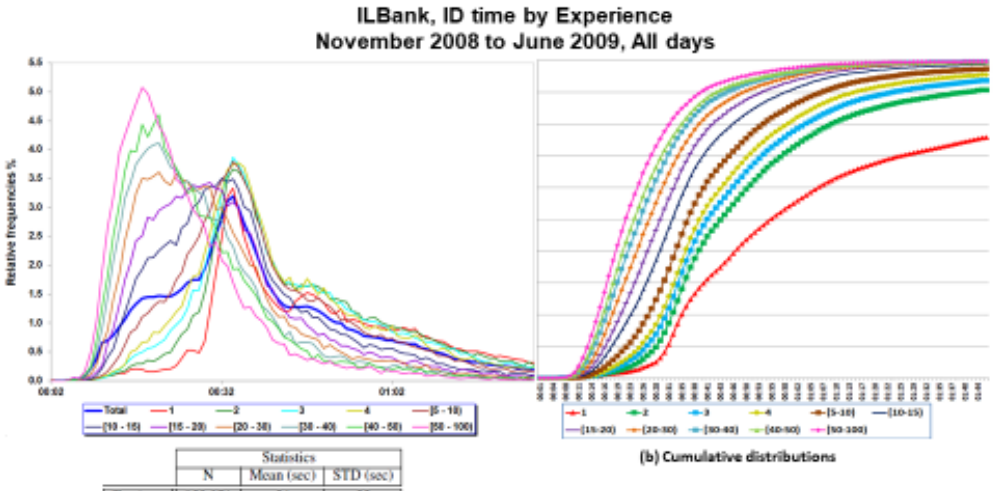
In the **"Options"** tab, under the **"Convert to"** select **"Frequencies"** and select chart type **"Polygon"**.

Click **"OK"**.



We observe that the distribution of ED occupancy (number of patients in the ED) at each time of day t, (t = 0, 1, ... , 23) seems normal, with mean and variance that vary over time (these statements were confirmed by the appropriate statistical tests, see Example 5.4 below). Using any of the 2 figures above, we identify three main patterns that constitute a mixture of the distribution in Example 5.1: (1) From 02:00 until 09:00, where the average number of patients is around 15 (ED lightly-loaded at night and early-morning); (2) From 12:00 until 22:00, where the average number of patients is about 35 (heavily-loaded from midday till late evening); and (3) The rest of the day (09:00–12:00, 22:00–02:00), when the distribution shifts from light- to heavy-loading (morning) and vice versa (evening-night).
This will be further observed in the next Example 5.3.

***Remark**: The idea behind the beautiful graph above is also relevant in other circumstances. Here is another example that arises from analyzing service durations in the answering machine of a call center (IVR or VRU).*



	Statistics		
	N	Mean (sec)	STD (sec)
Total	852,371	54	55
1	267,988	83	73
2	100,213	59	51
3	61,922	53	45
4	43,263	49	39
[5-10)	112,323	43	36
[10-15)	54,397	36	24
[15-20)	35,188	33	21
[20-30)	45,668	30	19
[30-40)	28,669	28	18
[40-50)	19,461	27	18
[50-100)	43,779	24	15

- הזדהות באמצעות קוד סודי ות.ז
- בחירת חשבון (מתוך רשימת חשבונות במידה וקיימים)

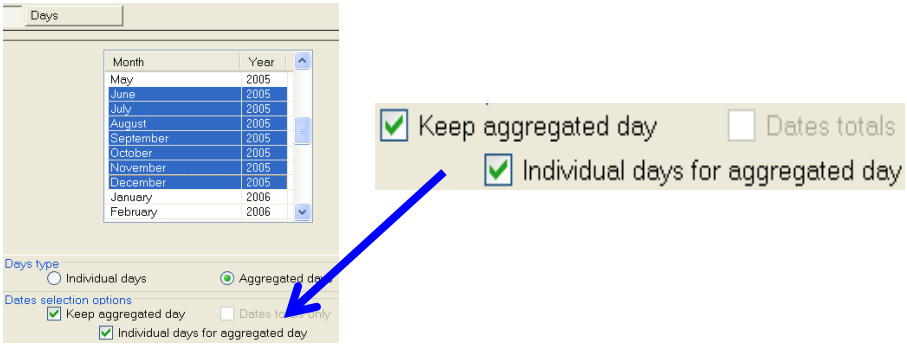
The histograms on the left depict experience of customers (after 1 use, 2 uses, [5-10] uses, …, [50-100] uses): the more experience accumulated the shorter is the duration. Thus, with experience, histograms move to the left (and, observing the right graph above, cumulative distribution functions move up – indeed, these durations are stochastically-ordered).

The IVR example is taken from the Technion [MSc thesis of Nitzan Carmeli](), Figures 6.11-6.12, which is entitled "Modeling and Analyzing IVR Systems, as a Special Case of Self-services."

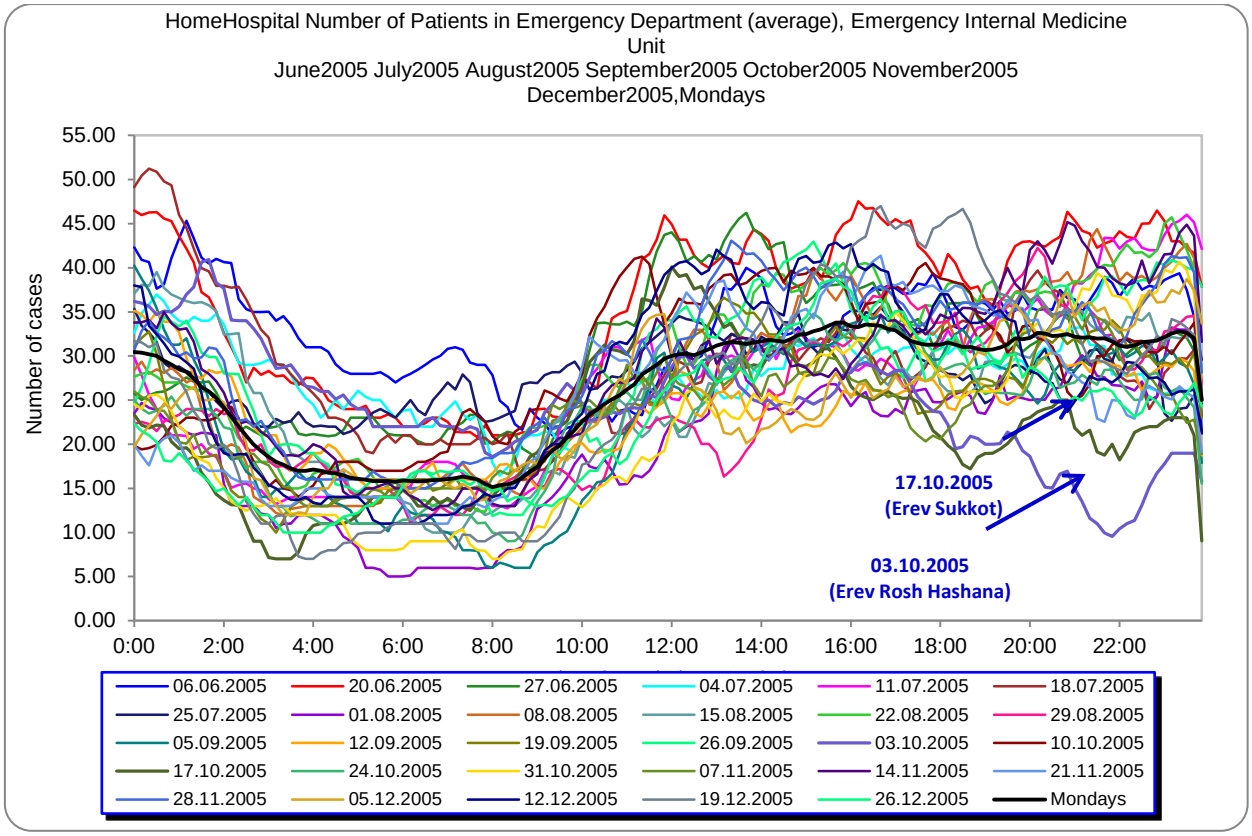
<u>Example 5.3:</u> Number of patients in Internal ED (Occupancy) - Average per 10-minute intervals, only on Mondays during 2005

Return to the **"Statistical Models (Summaries)"** window.
Click the **"New Model"** button. Select **"Time Series"**, then **"Intraday"**.
In the **"Variable"** tab, select **"Number of Patients in Emergency Department (average)"**.
In the **"Select Categories"** tab, select **"Emergency Internal Medicine Unit"**.
Click the **"Dates->"** button.
Mark **"Individual days for aggregated day"*.**
Select months from **June 2005** to **December 2005**.

*See Appendix E for explanations and examples of how to design a sample, in particular definitions of **"Individual days for aggregated day", "Dates totals only", "Add dates total", "Aggregated days"** (which covers all options).



Open the **"Days"** tab and select **"Mondays"**.
Click **"OK"**.



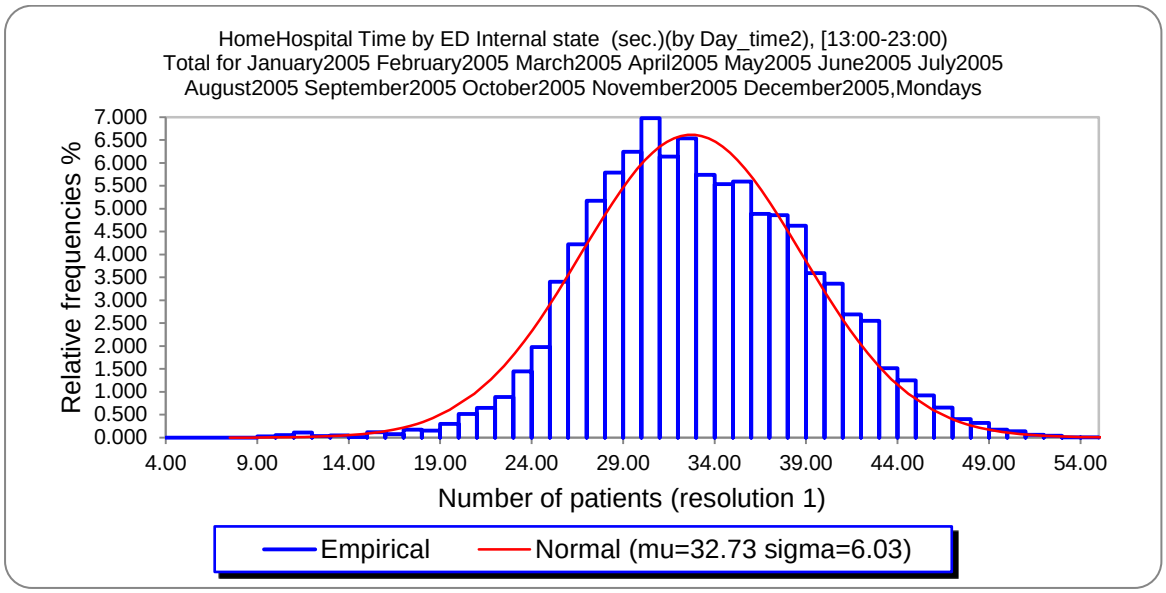
Note the drop in the evenings of 17/10/2005 and 03/10/2005: these are evenings before two Jewish holidays (erev Sukkot, erev Rosh Hashana), hence patients go home if at all possible. (One can identify such special days via **View->Calendar-> Individual days**->select month->click **Days** tab (see the **comments** column).

<u>Example 5.4</u>: Distribution of ED Occupancy (Time by ED Internal state (sec.) - Fitting a distribution during "evening" hours, on the Mondays of 2005

Return to the **"Statistical Models (Summaries)"** window.

Click the **"New Model"** button. Select **"Distributions"**, then **"Fitting"**. In the **"Variable"** tab, select **"Time by ED Internal state (sec.) (by Day_time2)"**. In the **"Options"** tab select **Normal** distribution. In the **"Select Categories"** tab, select **"[13:00–23:00)"**. Click the **"X Properties"** button, and select **"Upper Quantile (%)"** to be **100%**, if it is not already such).
Click the **"Dates->"** button.
Mark **"Dates totals only"**. Select months from **January 2005** to **December 2005**.

Open tab **"Days"** and select **"Mondays"**.
Click **"OK"**.



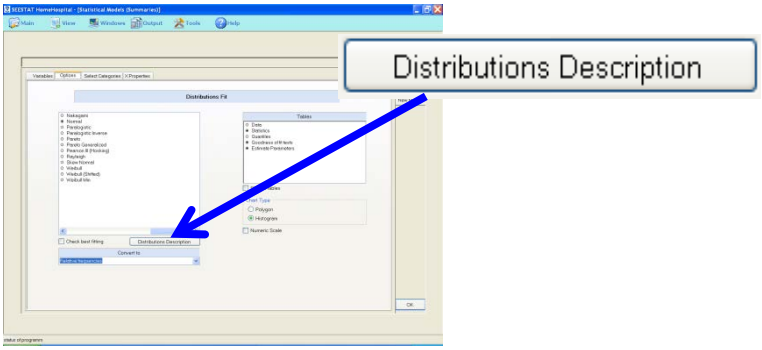
Statistics	
	Time by ED Internal state (sec.)
N	1836000
N(average per day)	36000
Mean	32.73
Standard Deviation	6.029
Variance	36.35
Median	32

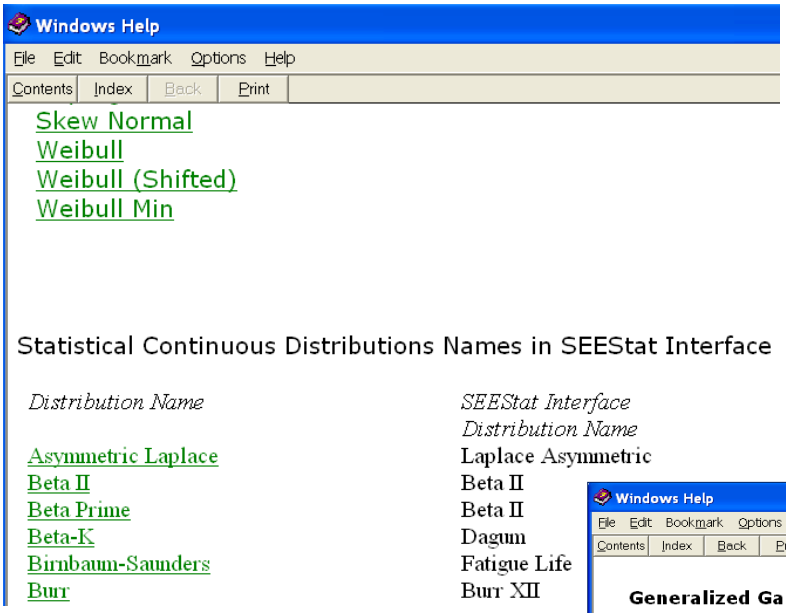
Parameters for Normal Distribution	
Parameter	*Estimate*
mu	32.73
sigma	6.03
mean	32.73
std	6.029

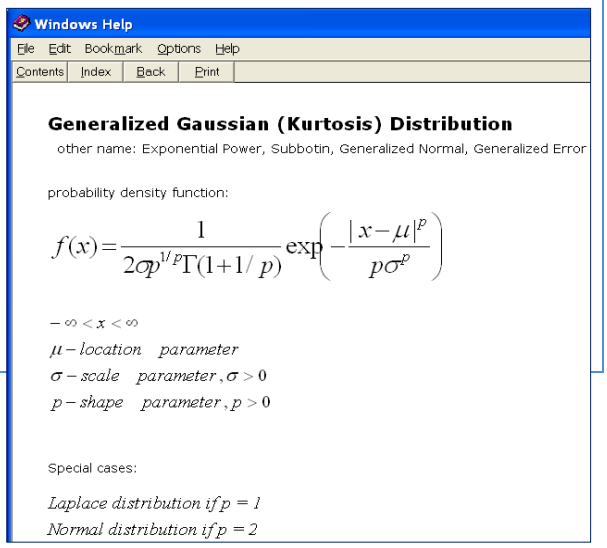
Goodness-of-Fit Tests for Normal Distribution			
Test	*Statistic*	*DF*	*p Value*
Residuals Std	0.0209		
Kolmogorov-Smirnov	0.0587		<.0001
Cramer-von Mises	801.9579		<.0001
Anderson-Darling	4532.0570		<.0001
Chi-Square	>1000	41	<.0001

Remark: In case you are unable to access these descriptions (e.g. faced with a Windows Help-window), please let me know.

***Note**: Formulas for 50 density functions, of continuous distributions, can be found under the "Options" tab, by clicking the "Distributions Description" button.*



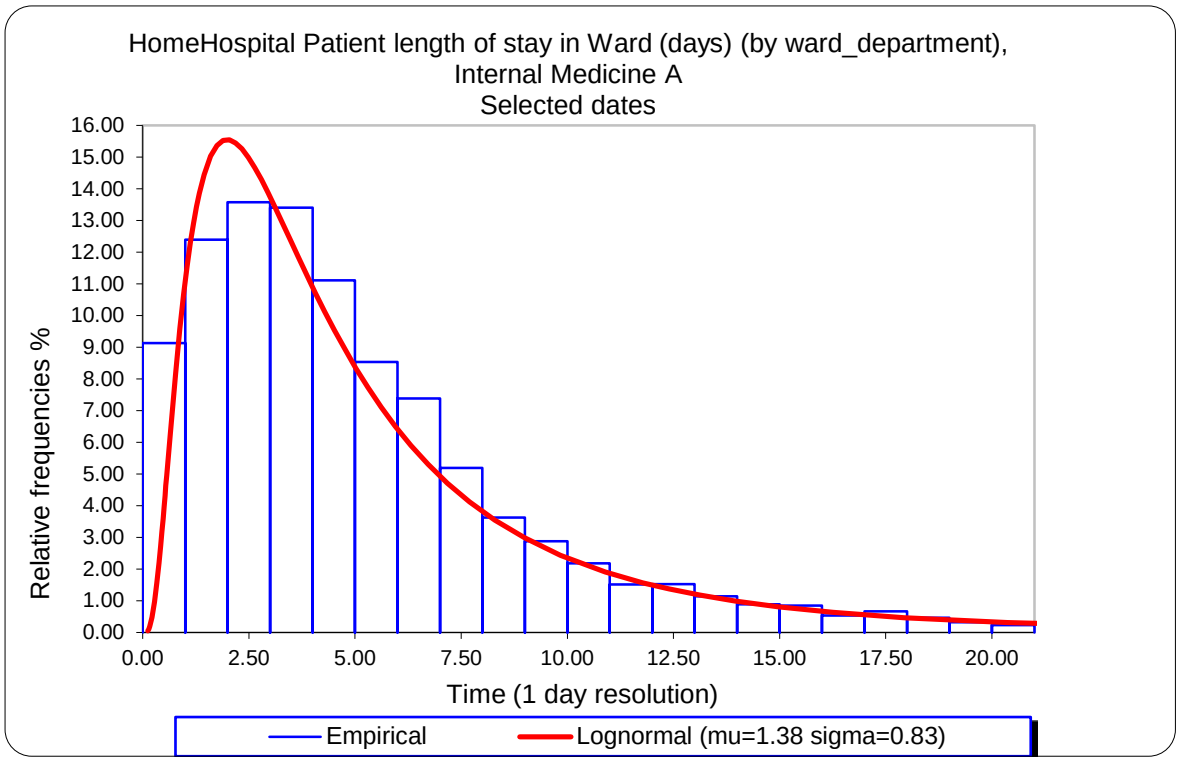




Part 6: Medical Wards

<u>Example 6.1:</u> Length-of-Stay (LOS) in Internal Wards (in days) – Distribution Fitting

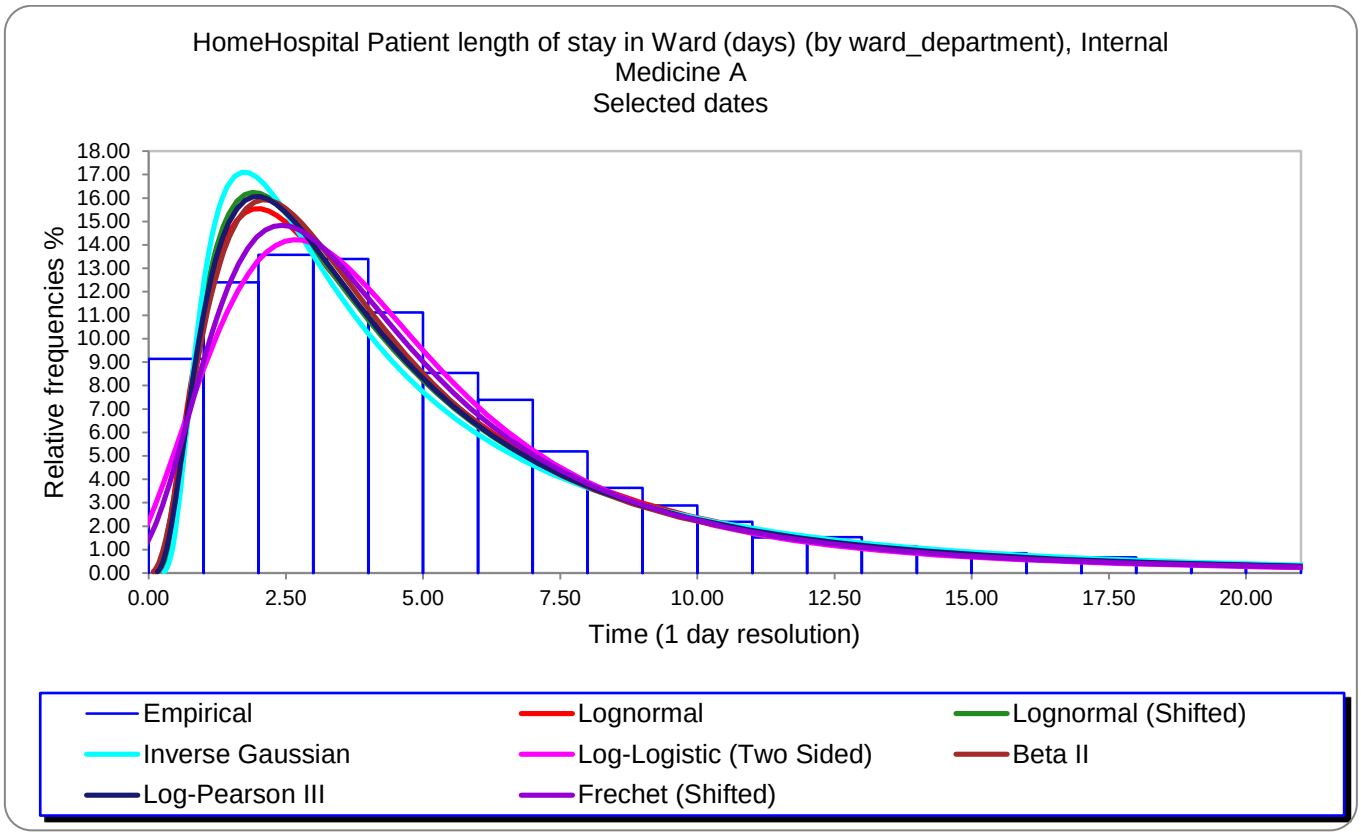
Return to the **"Statistical Models (Summaries)"** window.
Click the **"New Model"** button. Select **"Distributions"**, then **"Fitting"**.
In the **"Variable"** tab, select **"Patient length of stay in Ward (days) (by ward_department)"**.
In the **"Options"** tab select **LogNormal** distribution.
In the **"Options"** tab, under the **"Convert to"** select **"Relative frequencies".**
In the **"Select Categories"** tab, select **"Internal Medicine A"**.
In the **"X Properties"** tab, under **"Range to Display"** change the **Upper Quantile** limit to **97.5.** Click the **"Range to Compute"** button and choose "**Select Range"**, change the **Low Limit** to 1 and the **Upper Quantile** to 100.
Click the **"Dates->"** button.
Mark **"Dates totals only"**. Select months from **January 2004** to **October 2007** (using the shift key).
Open the **"Days"** tab and select **"All days"**.
Click **"OK"**.



When considering daily resolutions, the LogNormal distribution turns out to fit the LOS distribution well.
***Question**: Why is this a practically useful finding, beyond being scientifically intriguing? (There does not seem to be an explanation for the empirically-observed fact that*

LogNormal often captures duration of service processes: e.g. above, durations of phone calls when measured in seconds, and more.)

Exercise: *Repeat the above Example 6.1, but now click in* **"Options"** *the* **"Check best fitting"** *button; make sure to at least skim over the Excel Table that accompanies the Chart.*

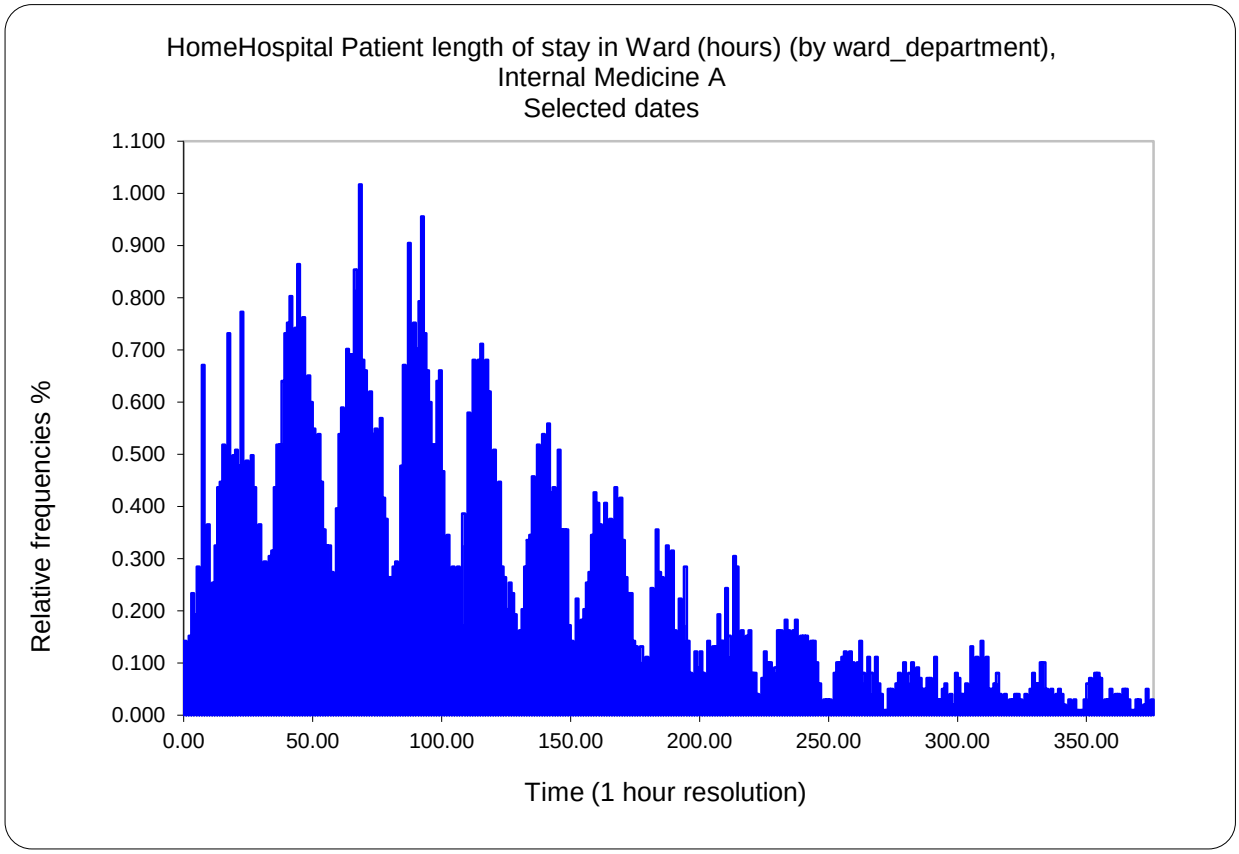


Example 6.2: *LOS in Internal Wards (in hours) – Protocol Mining*

Return to the **"Statistical Models (Summaries)"** window.
Click the **"New Model"** button. Select **"Distributions"**, then **"Estimates"**.
In the **"Variable"** tab, select **" Patient length of stay in Ward (hours) (by ward_department)"**.
In the **"Select Categories"** tab, select **"Internal Medicine A"**.
In the **"X Properties"** tab change the **Upper Quantile** limit to **95.**

Click the **"Dates->"** button.
Mark **"Dates totals only"**. Select months from **January 2004** to **October 2007**.
Open the **"Days"** tab and select **"All days"**.

Click **"OK"**.



In 1-hour resolution, we observe a completely different LOS distribution, with peaks that are periodically 24 hours apart. The reason for this phenomenon is the discharge protocol in hospital wards (which derives from the protocol of doctor rounds, after which they approve discharges, jointly with the protocols of nurse shifts – more on that momentarily). Thus, discharges are performed in "batches" of patients and, hence, takes only a few hours. This results in a very low variance of the discharge time, as most patients are released between 3pm and 4pm (which we shall now discover in Example 6.3).

Example 6.3: ***Patient Discharges from Ward - Intraday time series***

Return to the **"Statistical Models (Summaries)"** window; click **"New Model"**, select **"Time Series"** and **"Intraday"**.

In the **"Variables"** tab, select **"Patient Discharges from Ward"**.

In the **"Select Categories"** tab, select

"Department of Internal Medicine",

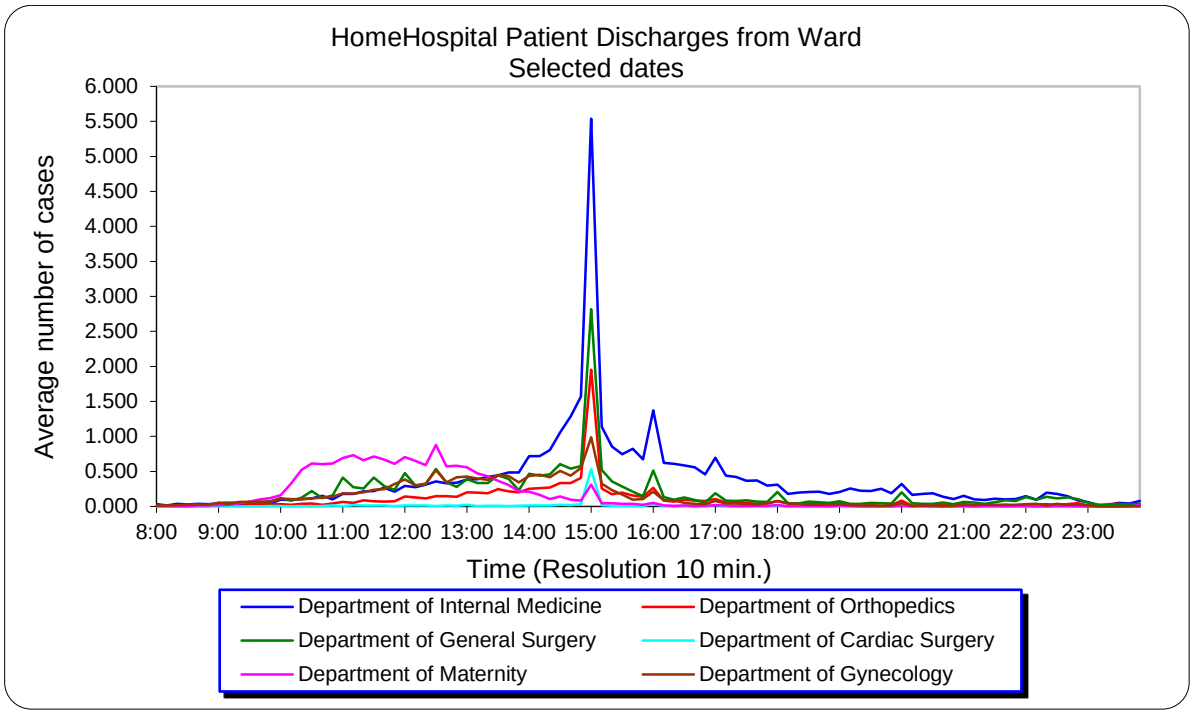
"Department of Orthopedics",

"Department of General Surgery",

"Department of Cardiac Surgery",

"Department of Maternity",
"Department of Gynecology".

Open the **"X Properties"** tab and change the low limit to **08:00.**
Click the **"Dates->"** button. Mark **"Dates totals only"** and select months from **January 2004 to October 2007**. Open the **"Days"** tab and select **"Weekdays"**.
Click **"OK"**, and observe the clear large peak at 15:00 (with some much smaller later peaks that are 1-hour apart from each other).



Example 6.4: Comments towards data-based research (of graduate students)

(Note: *While there are many web-links in this page, the text itself is self-contained without detours to them.)*

Some of our hospital EDA examples are taken from the paper **On Patient Flow in Hospitals: A Data-Based Queueing-Science Perspective**, published ***(***online***)*** in ***Stochastic Systems, Vol. 5 No. 1, 146-194, 2015***, and coauthored by Mor Armony, Shlomi Israelit (manager of the ED in Rambam Hospital, at the time of writing), Avishai Mandelbaum, Yariv Marmor, Yulia Tseytlin, and Galit Yom-Tov.

The paper builds on theses of three coauthors, who were AM's graduate students at Technion IE&M when the writing started:
- Yariv's PhD: "Emergency-Departments Simulation in Support of Service-Engineering: Staffing, Design, and Real-Time Tracking."
- Yulia's MSc: "Queueing Systems with Heterogeneous Servers: On Fair Routing of Patients in Emergency Departments."
- Galit's PhD: "Queues in Hospitals: Queueing Networks with ReEntering Customers in the QED Regime (QED = Quality- and Efficiency-Driven)."

The Patient-Flow paper uses the HomeHospital database, it focuses on the hospital sub-network <Emergency-Department → Internal Wards>, and it consolidates insights from the three theses to offer a data-based queueing-science perspective for patient flow in hospitals.

The paper is accessible in the following ***link.*** Note in particular Figure 1, which is a livelier rendering of the flow-chart at the beginning of Section HomeHospital Data; you can access its animation (we call it SEEnimation) in YouTube, via the link that is provided in the captions of Figure 1**.**
We also recommend that you read Section 6.2, which provides "Some concluding comments on data-based research—A great opportunity but no less of a challenge," followed by the APPENDIX: A MODEL FOR OR/AP DATA-BASED RESEARCH. This Appendix includes further information on SEELab, SEEStat and Reproducible Research.

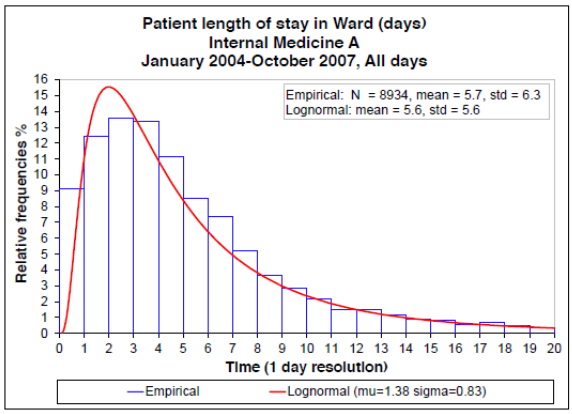
The Patient-Flow paper also has a long extended version, for those who would like to delve deeper into this exciting and important research subject. Indeed, and in addition to the above paper, all three theses also culminated in papers that were published in leading research journals and, as such, contributed to advancing knowledge-frontiers.
(Links to these papers, and many more, appear in AM's Service Engineering website, in its References menu.)

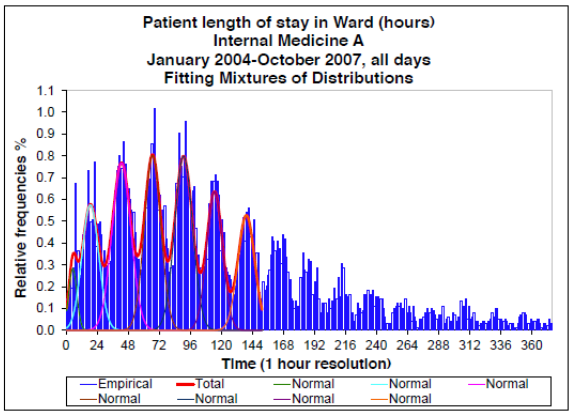
A last comment on **Reproducible Research**: to the extent possible, a data-based research paper must enable its readers to reproduce its data analysis. This serves 2 goals: first, verification of the results and, second, opening up the possibility of taking the research to its next phase/level.
As an example, here is a link, with material that guides readers on how to reproduce (most of) the analysis in the "On Patient Flow…" paper. Our last example below is extracted from the second document in that link. It ought to demonstrate how SEEStat output/EDA can be taken one small step forward, thus creating an Excel graph, which appears in the paper and is "worth at least 1000 words".

6.4.1 "EDA: LOS - a story of multiple time scales" – taking SEEStat forward

The following was extracted, with minor format changes, from the following document: EDA via SEEStat 3.0 to Reproduce **"Patient Flow in Hospitals: A Data–Based Queueing–Science Perspective"** (Link)

Reproducing Fig 9: LOS distribution of IW A in two time-scales: daily and hourly



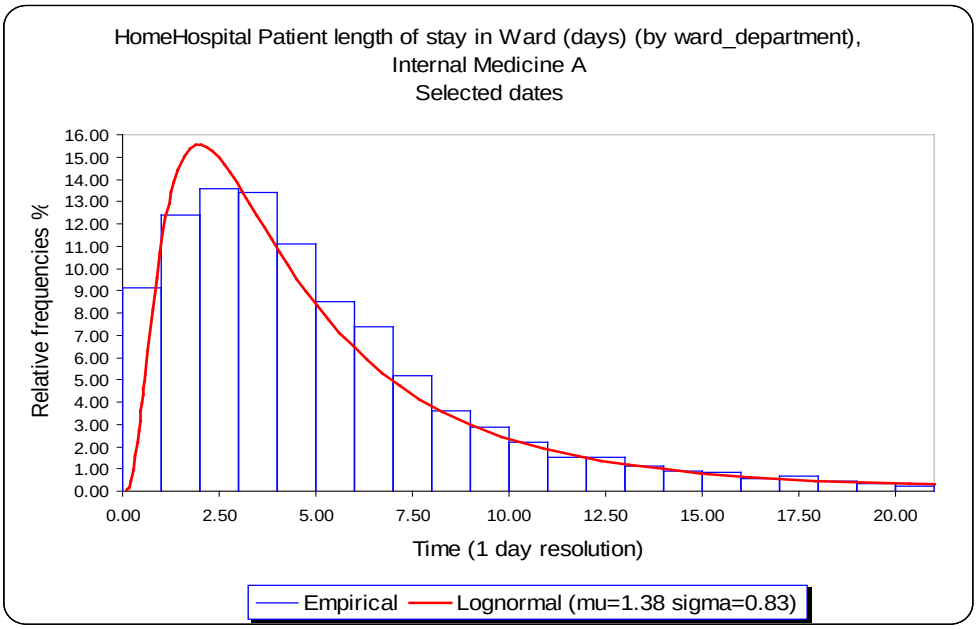


Reproducing steps:

1. Creating a chart:

   Click **"Main"** –> **"Statistical Models (Summaries) "**. Select **"Distribution"**, then **"Fitting"**. From the variables list, select **"Patient length of stay in Ward (days) (by ward_department) "**. In the **"Options"** tab, select **"Lognormal"** distribution. In the **"Select Categories"** tab, select **"Internal Medicine A"**. Open the **" Properties"** tab. Select resolution **1** day; range to display: low limit – minimal value, upper limit–**97.5%,** and range to compute: low limit **1**, upper limit **100%.** Click the **"Dates –> "**button. Select **Dates totals only** and all months from **January 2004** to **October 2007**. Open tab **"Days"** and select **"All days"**. Click **"OK"**.

Original SEEStat chart:



Original SEEStat table:

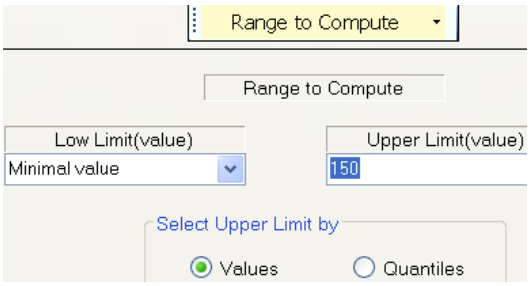
Statistics	
	Patient length of stay in Ward (days) (by ward_department)
N	8934
N(average per day)	6.381428571
Mean	5.665
Standard Deviation	6.284
Variance	39.49
Median	4
Minimum	1
Maximum	151
Skewness	6.026
Kurtosis	76.67
Standard Error Mean	0.0665
Interquartile Range	5
Mean Absolute Deviation	3.688
Median Absolute Deviation(MAD)	2
Coefficient of Variation(CV) (%)	110.92
L-moment 2 (half of Gini's Mean Difference)	2.552
L-Skewness	0.411
L-Kurtosis	0.271
Coefficient of L-variation(L-CV)(%) (Gini's Coefficient)	45.05

Parameters for Lognormal Distribution	
Parameter	*Estimate*
mu	1.38
sigma	0.83
mean	5.593
std	5.579

2. <u>Designing the original SEEStat chart</u>:

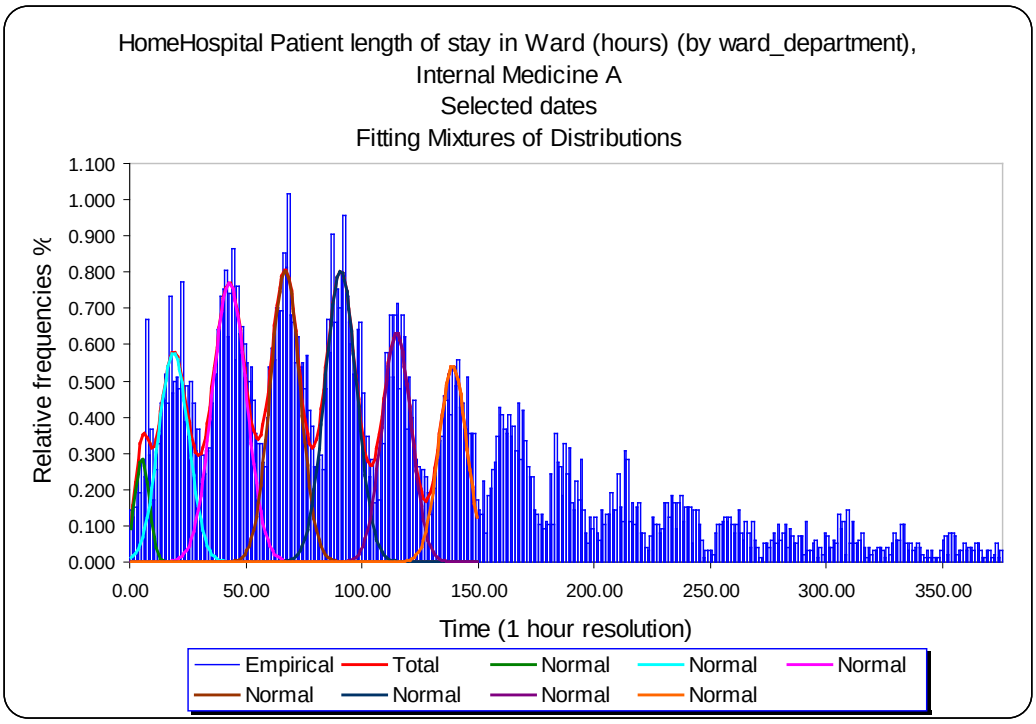
   Change in the chart title *Selected dates* to January 2004–October 2007, All days.

3. Click **"New Model"** button. Select **"Distribution"**, then **"Mixture fitting"**. From the variables list, select **"Patient length of stay in Ward (hours) (by ward_department)"**. In the **"Options"** tab, select **"Normal"** distribution. Set the number of mixture components to **7**. In the **"Select Categories"** tab, select **"Internal Medicine A"**. Open the **"Properties"** tab, select resolution **1** hour, click on **"Range to Compute"** button, then **"Select Range"**, mark **Values** and set upper limit **150** = 150 hours, in **"Range to Display"**: low limit – **minimal value**, upper limit – **95%**.



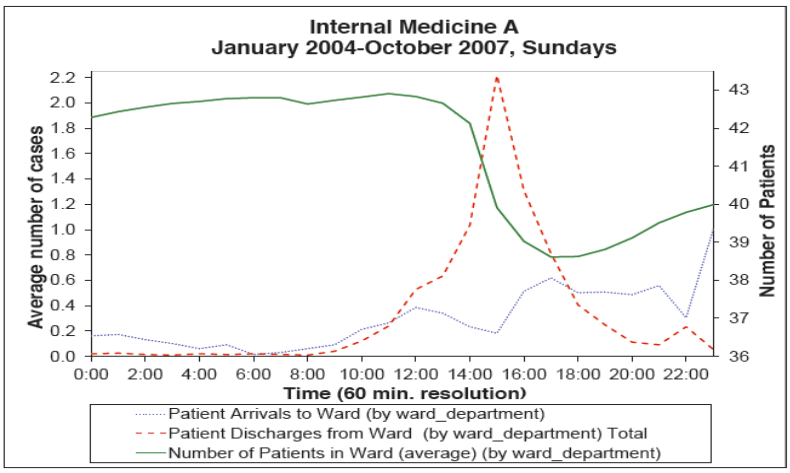
Click the **"Dates –>"** button. Select **"Dates totals only"** and all months from **January 2004** to **October 2007**. Open tab **"Days"** and select **"All days"**. Click **"OK"**.

Original SEEStat chart:



4. <u>Design the original SEEStat chart</u> – this entails the following 2 changes to the original graph:
   i. Change in the chart title *Selected dates* to *January 2004 – October 2007, All days.*
   ii. Format horizontal axis: set major unit *24*, decimal places *0.*

Reproducing Fig 10: Arrivals, departures, and average number of patients in Internal wards by hour of day

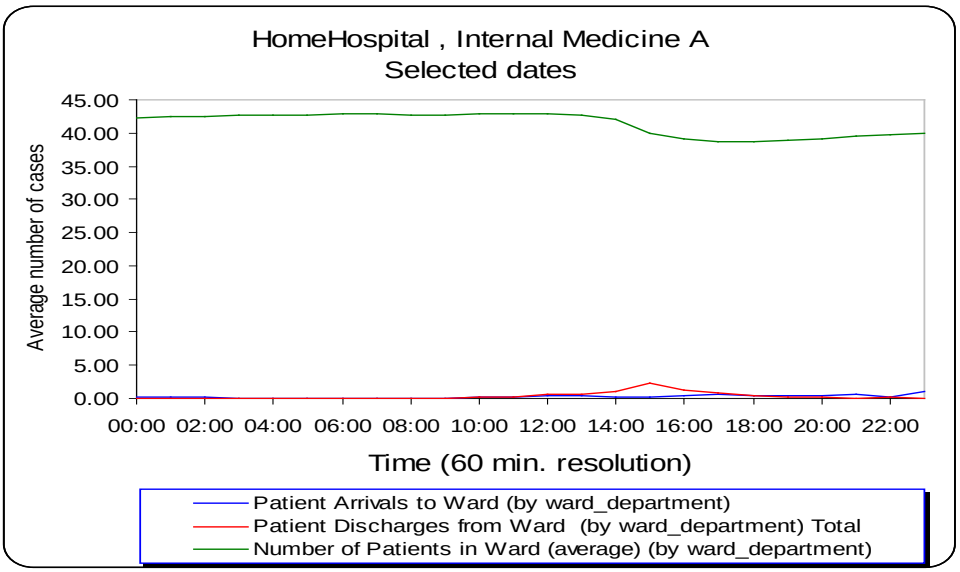


Reproducing steps:

1. Creating chart:

   Click **"Main"** –> **"Statistical Models (Summaries) ".** Select **"Time Series"**, then **"Intraday"**.
   From the variables list, select **"Patient Arrivals to Ward (by_ward_department) ", "Patient Discharges from Ward (by ward_department) "** and **"Number of Patients in Ward (average) (by ward_department) ".** In the **"Select Categories"** tab, select **"Internal Medicine A"**. Open the **"Properties"** tab. Select resolution 60:00 = 1 hour.
   Click the **"Dates –>"** button. Select **"Dates totals only"** and all months from **January 2004** to **October 2007**. Open tab **Days** and select **Sundays.**
   Click **"OK"**.

Original SEEStat chart:

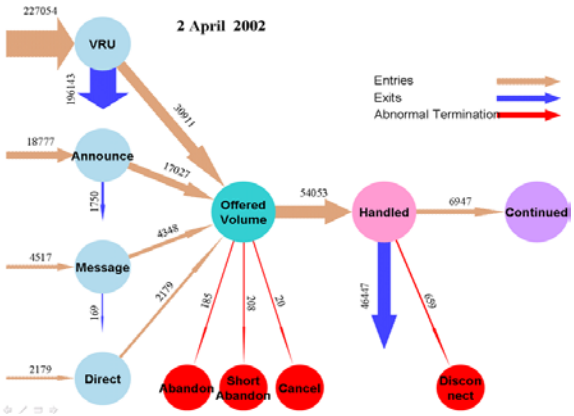


2. Design the original SEEStat chart - this entails the following 3 changes to the original graph:
   a. Add and format a secondary vertical axis for *Number of Patients in Ward (average) (by ward_department).*
   b. Format primary vertical axis, in order to change the maximal value.
   c. Change the chart title from "*Selected dates*" to "*January 2004 – October 2007, Sundays*".

The Excel process of carrying out these changes was extracted from the above paper as well, and it is reproduced in Appendix D for your convenience.

Appendix A: On SEEnimations (Data-Animations)

We started the tutorial with a simple explanation of the data structure of USBank call center – one that used the following simple daily flow-chart (April 2, 2002):

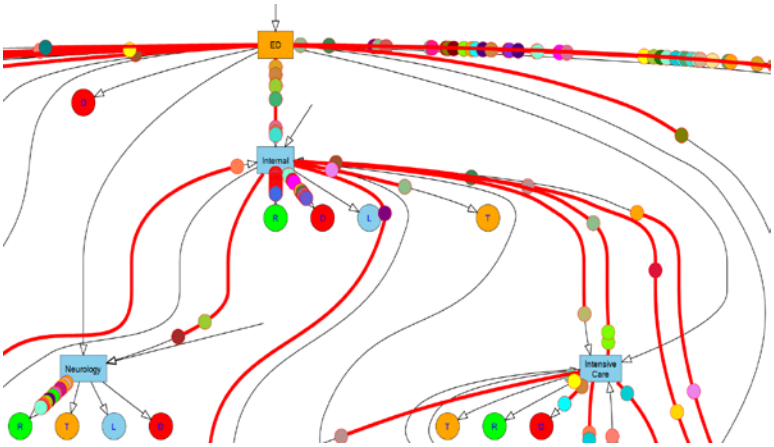


Creating such flow-charts is very easy in SEEStat: Leave SEEStat, restart it, select in ***"Main"*** *the option* ***"Daily Reports",*** *chose again the database* ***USBank,*** *select both options* ***"Flow chart"*** *and* ***"Report"****, then chose in* ***"Dates"*** *the above* ***date*** *(or any other date, or aggregated-days and then a month) as you learned during our tutorial. Your flow-chart will then unfold as a PowerPoint file; and, in parallel, SEEStat will create a numerical summary as an Excel table.*

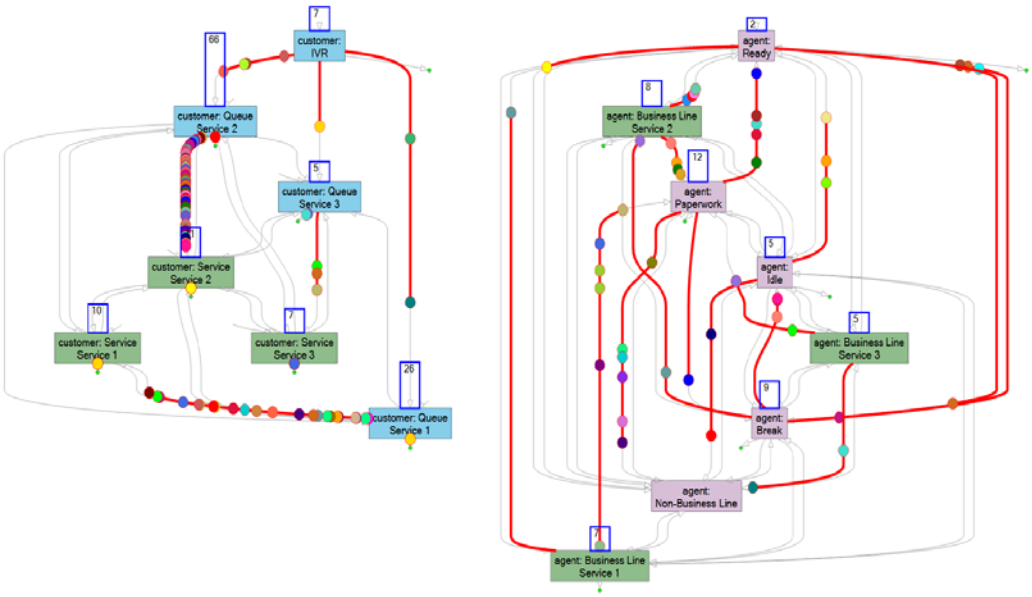
But this, and in fact your whole tutorial experience, is only the tip of the iceberg – SEEStat has several "relatives" that can take you far beyond basic EDA. For example, consider ***SEEGraph***, which creates semi-automatically data-animations – we call these ***SEEnimations*** (and you can find many of them by searching <seenimations> in YouTube).

As an appetizer, we now take you through several examples of SEEnimations:

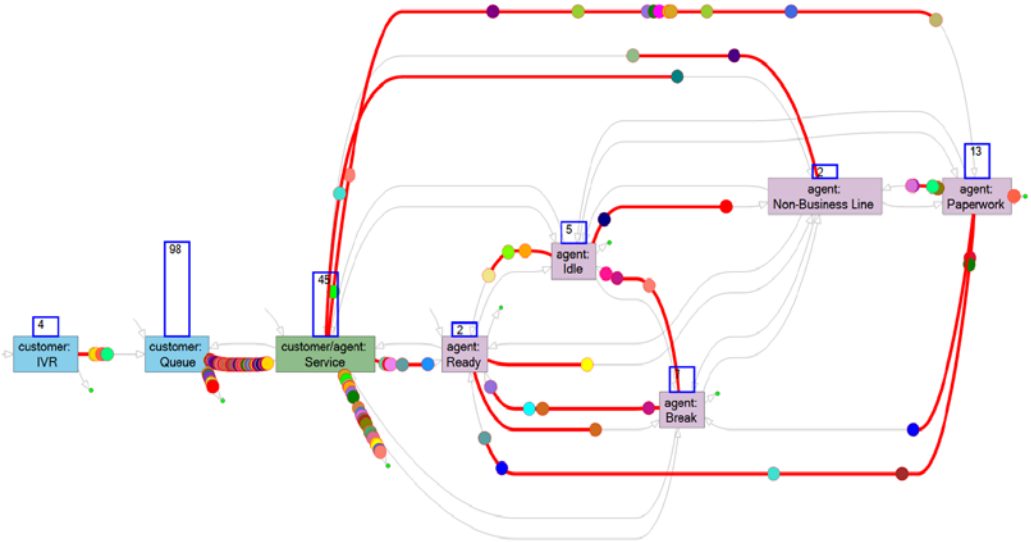
1. Potpourri of data-animations from hospitals and call centers, where you can observe of hospitals (patients) and call centers (customers): link



2. Call Center: Customers Network and Agents Network, separated but viewed together: link



3. Redoing Example 2, but now customers and agents become a single Resource Network: link.



Examples 2 and 3 are taken from the following BSF research proposal:
Mandelbaum A., Armony M. and Momcilovic P. **Data-Based Models of Resource-Driven Activity Networks.** (PDF-P47KB)
It sets the foundation for Chapter 4 in the following PhD thesis proposal:
Carmeli N. **Data-Based Resource-View of Service Networks: Performance Analysis, Delay Prediction and Asymptotics.** Ph.D. Research Proposal, Technion, February 2017. (PDF-2.18MB)

Final comments: Again, the above is only the tip of the iceberg. Experiencing SEEGraph, and learning how to create your own SEEnimations, are the subject of several separate tutorials. I shall hopefully be able to teach the first of them (Understanding SEEGraphs and SEEnimations) sooner than later.

Appendix B: Mixture- and Distribution-Fitting

Mixture Fitting

The mixture distribution function is given by:

$$p(x) = \sum_{j=1}^{k} w_j \varphi(x;\theta_j); \qquad \sum_{j=1}^{k} w_j = 1; \qquad w_j > 0$$

k – given number of components of the mixture
φ – given distribution function (Normal, Lognormal, Gamma, Exponential, Weibull or Invers
Gaussian)
θ_j – distribution parameters (shape, scale or location parameters)
w_j – weight coefficient or mixing proportions

Task: Find optimal w_j and θ_j for sample X and given k and φ.

Solution:

- Specify initial values for w_j by using grid.
- For each point in a grid calculate mean, standard deviation (specify initial values for θ_j) and objective function (sum of squares of difference between empirical and fitted cumulative distribution functions).
- Select points with smallest values of objective function (best points).
- For each best point compute fitted distribution by using optimization algorithm.
- The best fit with smallest value of objective function is selected.

Distribution Fitting (50 distributions)

Algorithms of distribution fitting (SEEStat software):

1. *Maximum likelihood estimates.*
   (11 distributions)
2. *L-moments estimates* (linear combination of order statistics).
   (31 distributions)
3. *Maximization of likelihood function.*
   (7 distributions)
4. *Optimization of the objective function* (sum of squares of differences between empirical and fitted cumulative distribution functions)
   (1 distribution)

Appendix C: Smoothing References

Heft: Hazard estimation with flexible tails

[1] Charles Kooperberg, Charles J. Stone and Young K. Truong (1995). Hazard regression. *Journal of the American Statistical Association*, **90**, 78–94.

[2] Charles J. Stone, Mark Hansen, Charles Kooperberg, and Young K. Truong (1997). The use of polynomial splines and their tensor products in extended linear modeling (with discussion). *Annals of Statistics*, **25**, 1371–1470.

Loess: Local polynomial regression fitting

[3] Cleveland, E. Grosse and W.M. Shyu (1992). Local regression models. Chapter 8 of *Statistical Models in S*, eds J.M. Chambers and T.J. Hastie, Wadsworth & Brooks/Cole.

Muhaz: Estimate hazard function from right-censored data

[4] H.G. Mueller and J.L. Wang (March 1994). Hazard rate estimation under random censoring with varying kernels and bandwidths, *Biometrics*, **50**, 61–76.

[5] O. Gefeller and H. Dette (1992). Nearest neighbour kernel estimation of the hazard function from censored data, *J. Statist. Comput. Simul*., **43**, 93-101

[6] K.R. Hess, D.M. Serachitopol and B.W. Brown (1999). Hazard function estimators: A simulation study, *Statistics in Medicine*, **18** (22), 3075–3088.

Supsmu: Friedman's SuperSmoother

[7] J.H. Friedman (1984). SMART User's Guide. Laboratory for Computational Statistics, Stanford University Technical Report No. 1.

[8] J.H. Friedman (1984). A variable span scatterplot smoother. Laboratory for Computational Statistics, Stanford University Technical Report No. 5.

Pspline: Fit a polynomial smoothing spline of arbitrary order

[9] J.O. Ramsay, N. Heckman and B.W. Silverman (1997). Spline smoothing with model based penalties. *Behavior Research Methods, Instruments, & Computers*, **29** (1), 99–106.

Bspline: Fits a cubic smoothing spline to the supplied data

[10] J.M. Chambers and T.J. Hastie (1992). *Statistical Models in S*, Wadsworth & Brooks/Cole.

[11] P.J. Green and B.W. Silverman (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach.* Chapman and Hall.

[12] T.J. Hastie and R.J. Tibshirani (1990). *Generalized Additive Models.* Chapman and Hall.

dpill: Select a bandwidth for local linear regression

[13] D. Ruppert, S.J. Sheather and M.P. Wand (1995). An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association*, **90**, 1257–1270.

[14] M.P. Wand and M.C. Jones (1995). *Kernel Smoothing.* Chapman and Hall, London.

dpih: Select a histogram bin width

[15] D.W. Scott (1979). On optimal and data-based histograms. *Biometrika*, **66**, 605–610.

[16] S.J. Sheather and M.C. Jones (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society, Series B*, **53**, 683–690.
[17] M.P. Wand (1995). Data-based choice of histogram bin width. *University of New South Wales*, Australian Graduate School of Management Working Paper Series No. 95–011.

dpik: Select a bandwidth for kernel density estimation

[18] S.J. Sheather and M.C. Jones (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society, Series B*, **53**, 683–690.
[19] M.P. Wand and M.C. Jones (1995). *Kernel Smoothing.* Chapman and Hall, London.

locpoly: Estimate functions using local polynomials

[20] M.P. Wand and M.C. Jones (1995). *Kernel Smoothing.* Chapman and Hall, London.

density: Kernel Density Estimation

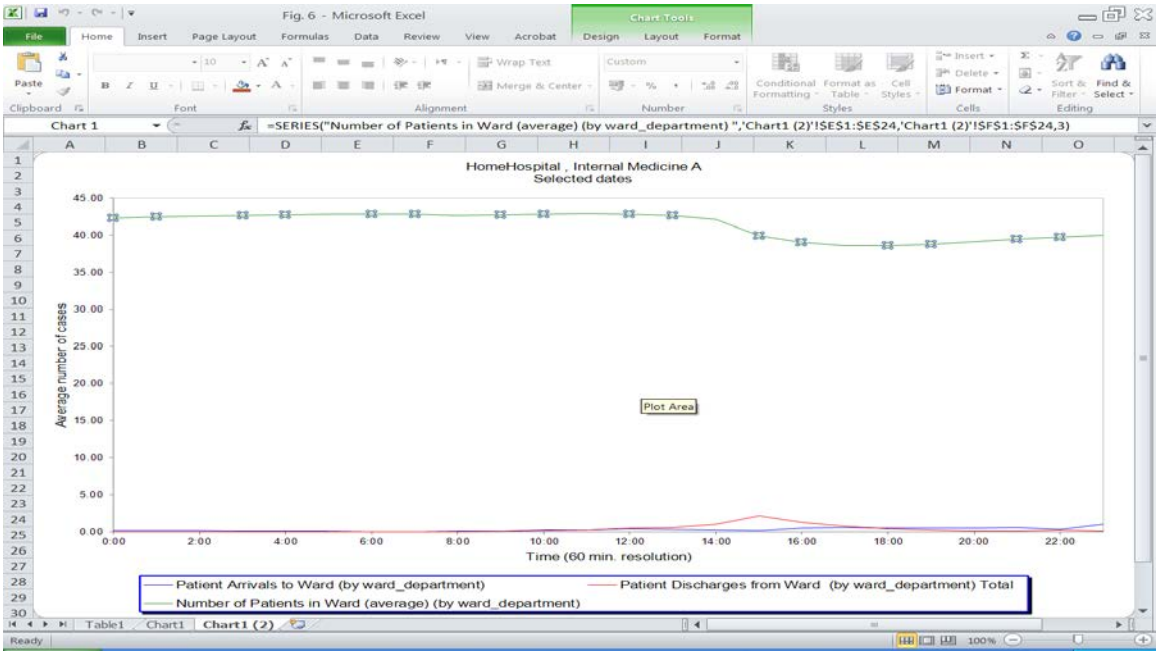
[21] R.A. Becker, J.M. Chambers and A.R. Wilks (1988). *The New S Language*. Wadsworth & Brooks/Cole (for S version).
[22] D.W. Scott (1992) *Multivariate Density Estimation. Theory, Practice and Visualization*. New York: Wiley.
[23] S.J. Sheather and M.C. Jones (1991). A reliable data-based bandwidth selection method for kernel density estimation. *J. Roy. Statist. Soc.* **B**, 683–690.
[24] B.W. Silverman (1986). *Density Estimation*. London: Chapman and Hall.
[25] W.N. Venables and B.D. Ripley (2002). *Modern Applied Statistics with S*. New York: Springer.

bandwidth: Bandwidth selectors for kernel density estimation

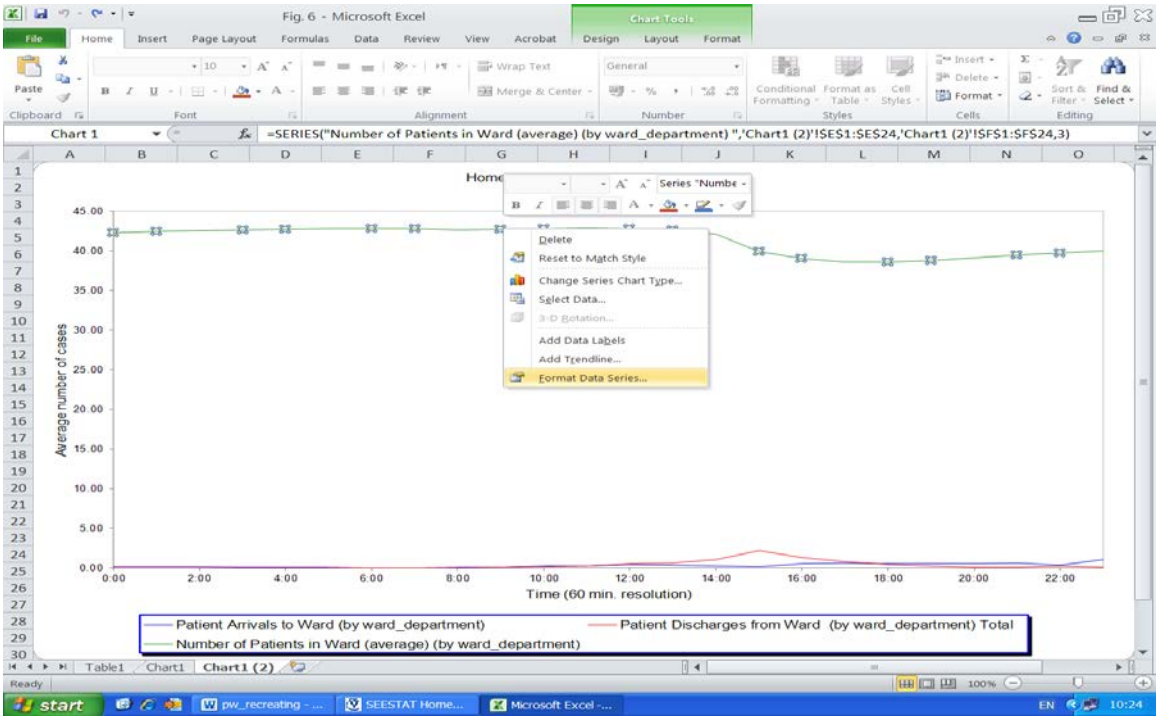
[26] D.W. Scott (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley.
[27] S.J. Sheather and M.C. Jones (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society series B*, **53**, 683–690.
[28] B.W. Silverman (1986). *Density Estimation*. London: Chapman and Hall.
[29] W.N. Venables and B.D. Ripley (2002). *Modern Applied Statistics with S*. Springer.
[30] Adrian W.Bowman and Adelchi Azzalini (1997). *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations*. Clarendon Press Oxford
[31] Jeffrey S.Simonoff (1996). *Smoothing Methods in Statistics*. Springer
[32] T.J. Hastie and R.J. Tibshirani (1990). *Generalized Additive Models*. Chapman & Hall/CRC.

Appendix D: Adding a Secondary Vertical Axis to an Excel Figure

1. Add a secondary vertical axis in the chart: right click on data series "Number of Patients in Ward (average) (by ward_department)" (green line)

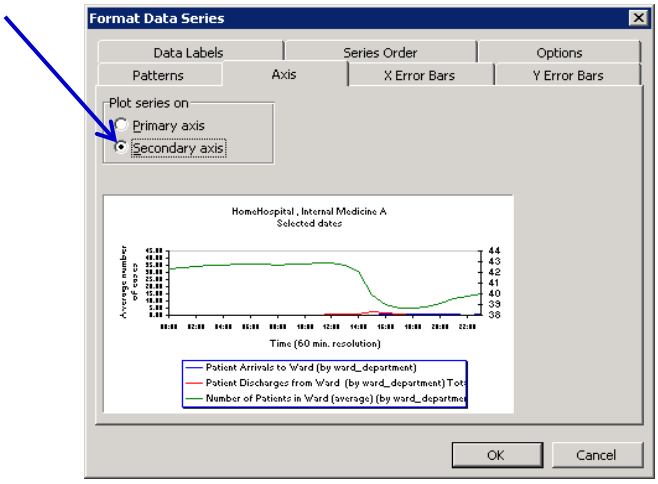


2. Select Format Data Series…

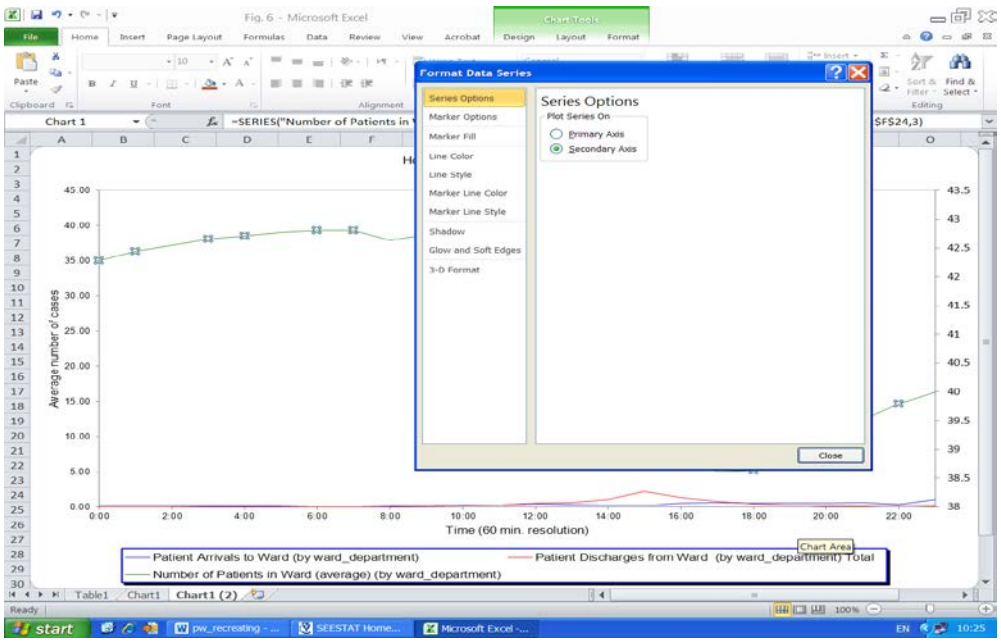


3. Select Secondary Axis

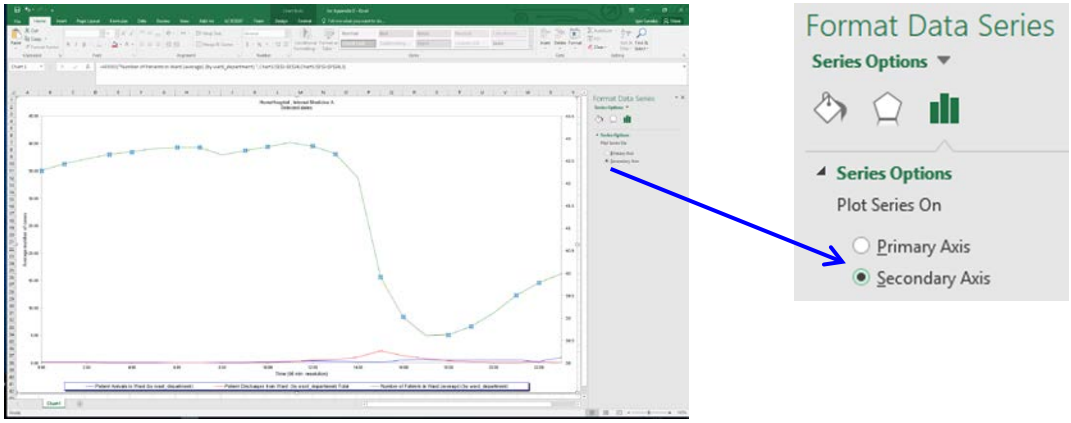
Excel 2003: Click on Axis tab and select plot series on Secondary axis.



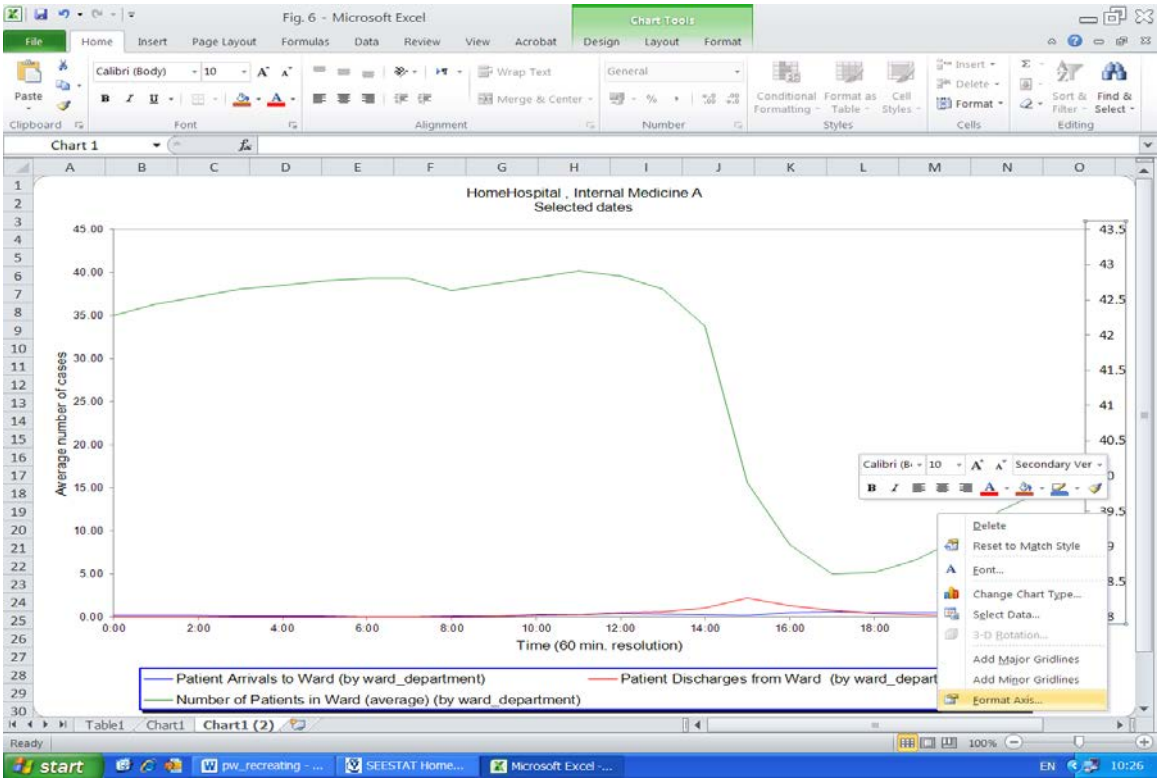
Excel 2010: in Series Options select Secondary Axis.



Excel 2016: in Series Options select Secondary Axis.

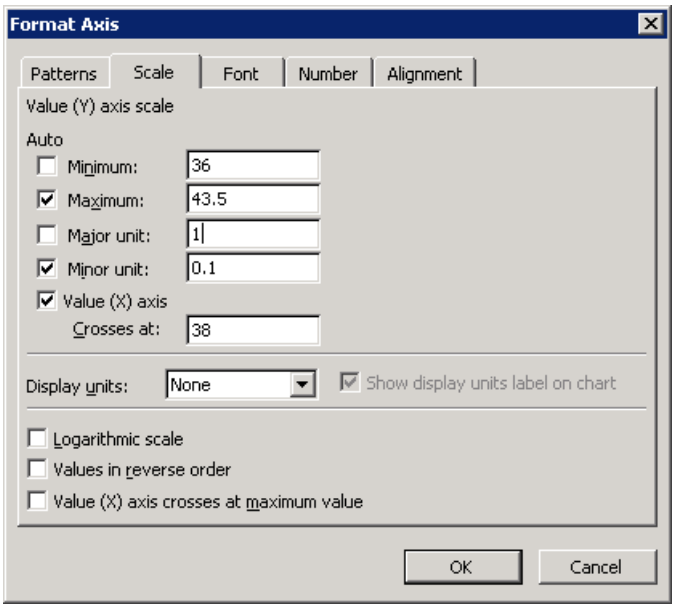


4. Right click on secondary axis and select Format Axis …



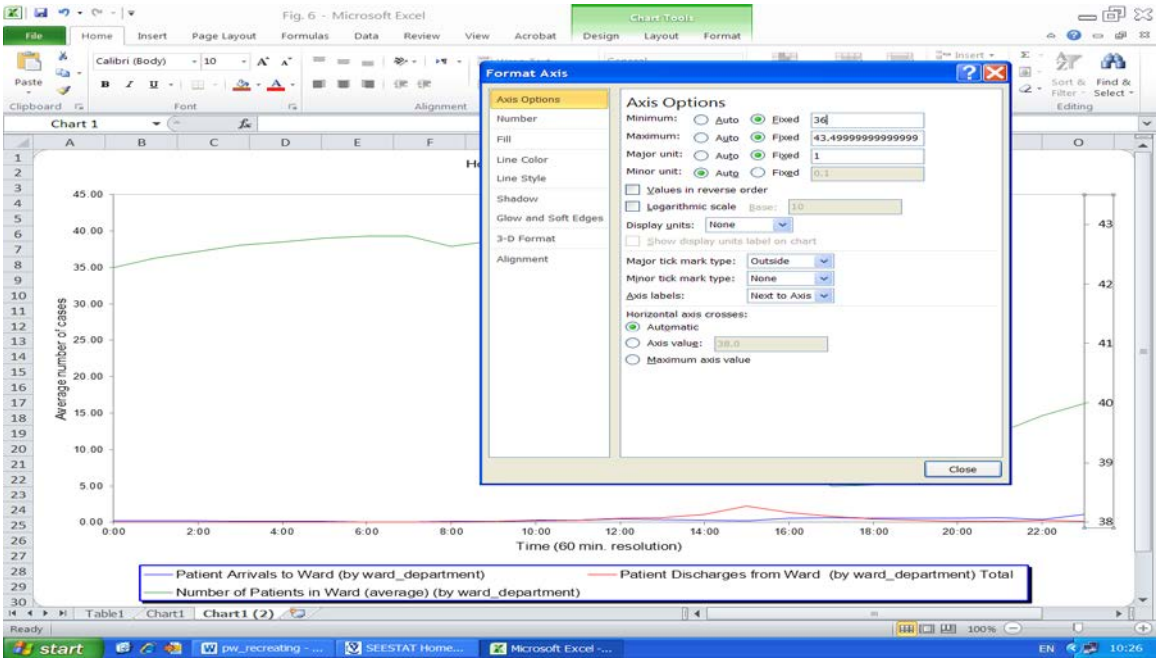
5. Fill in axis options minimum 36, maximum 43.5 major unit 1

Excel 2003: click on tab Scale and fill in parameters.

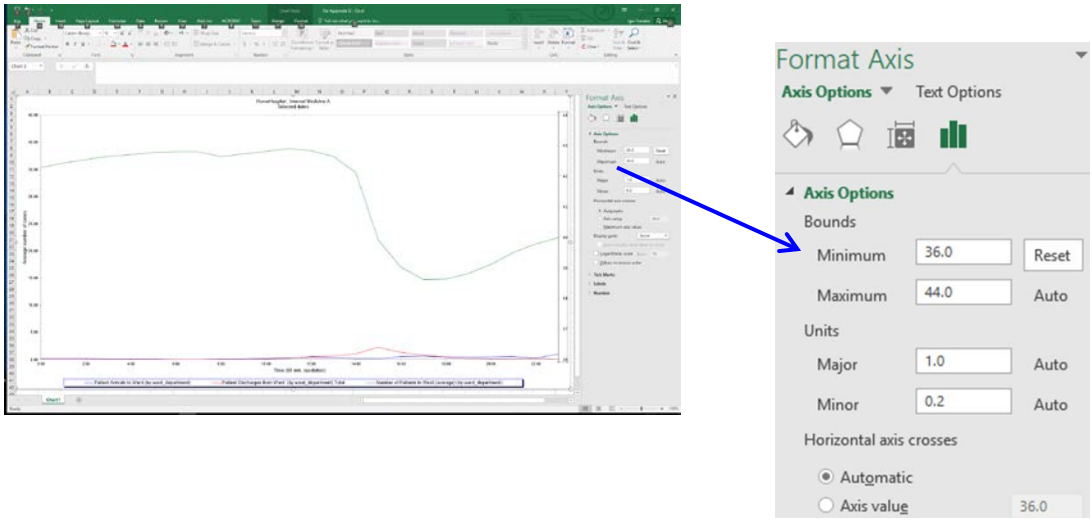


.

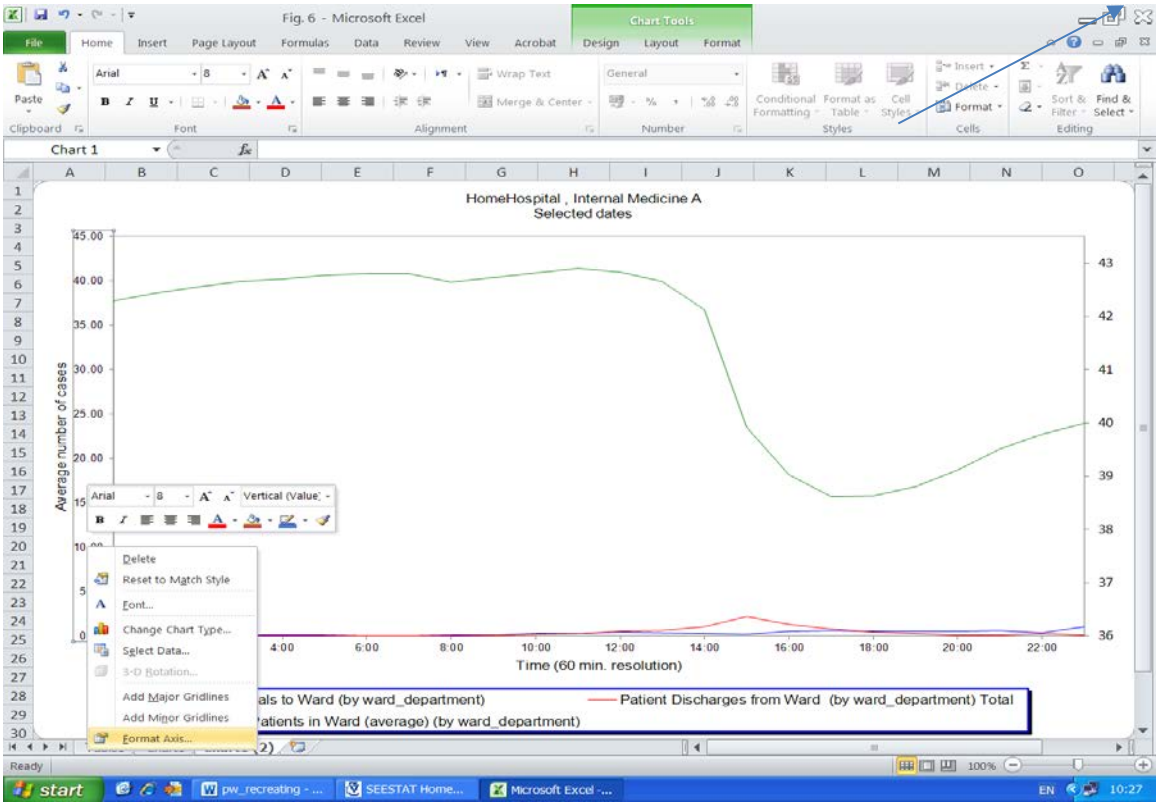
Excel 2010: select Axis Options and fill in parameters



Excel 2016: select Axis Options and fill in parameters.

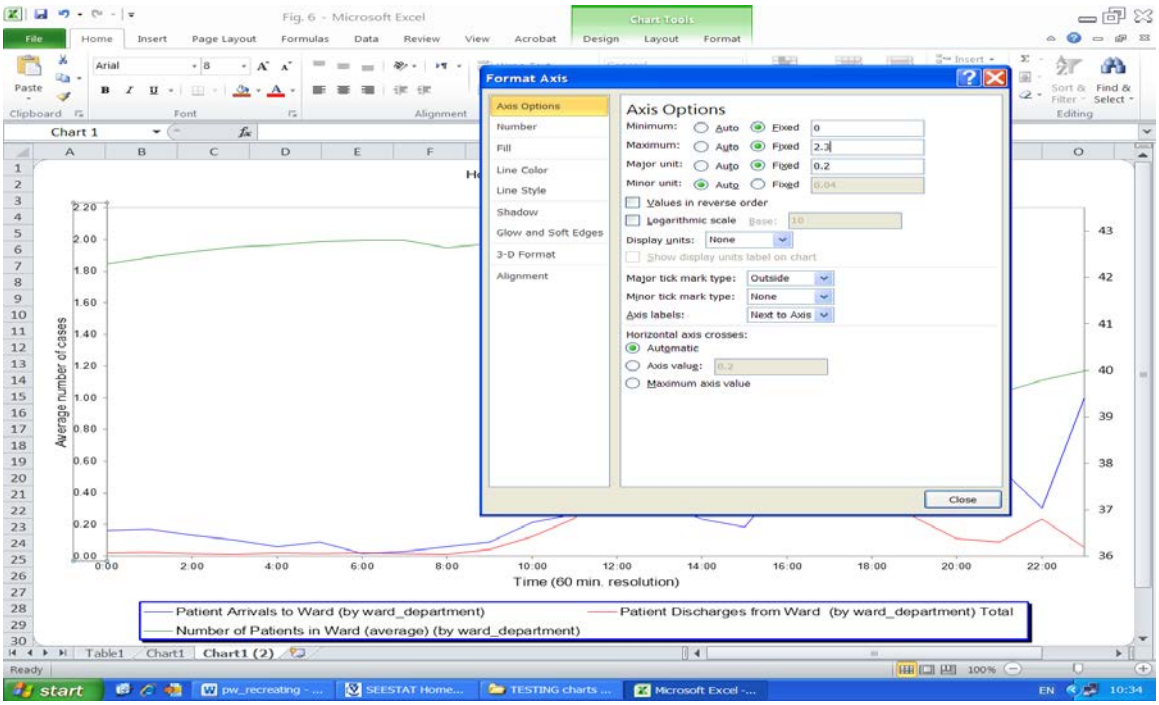


6. Right click on primary axis (left side) and select Format Axis …

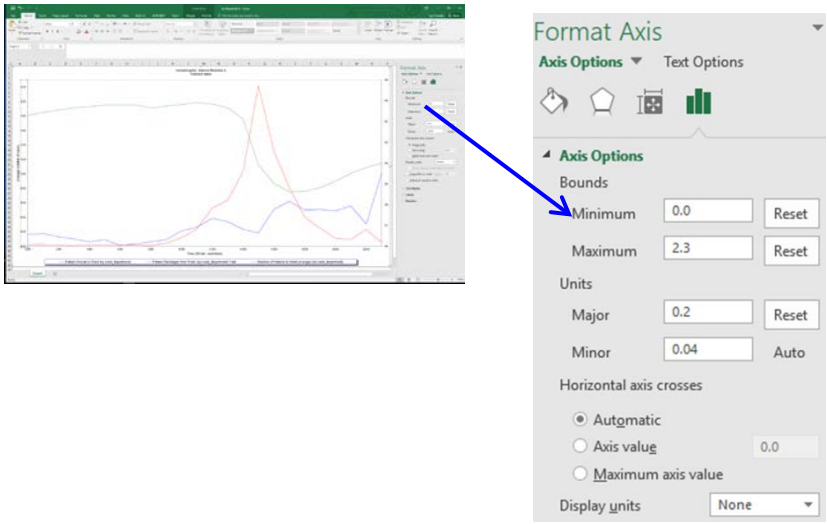


7. Fill in axis options minimum 0, maximum 2.3 major unit 0.2

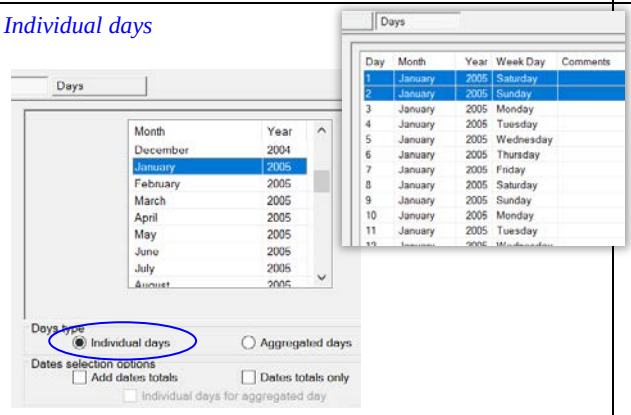
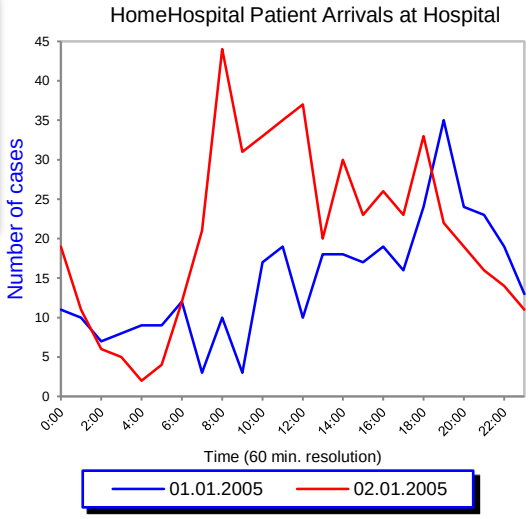
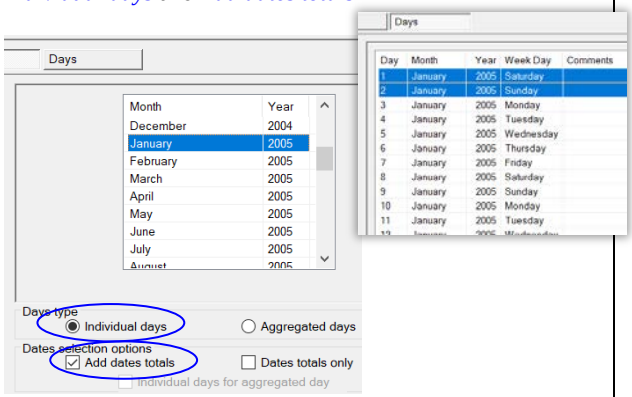
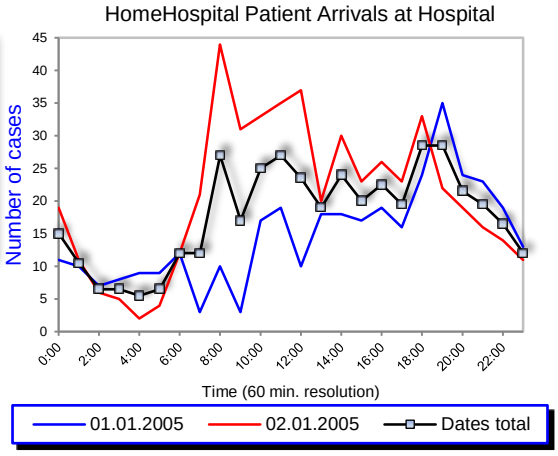
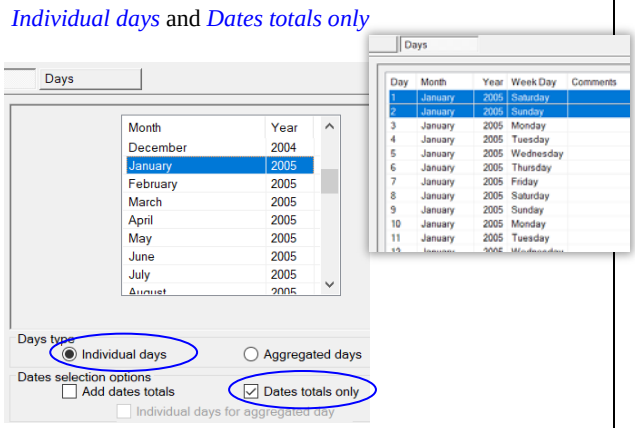
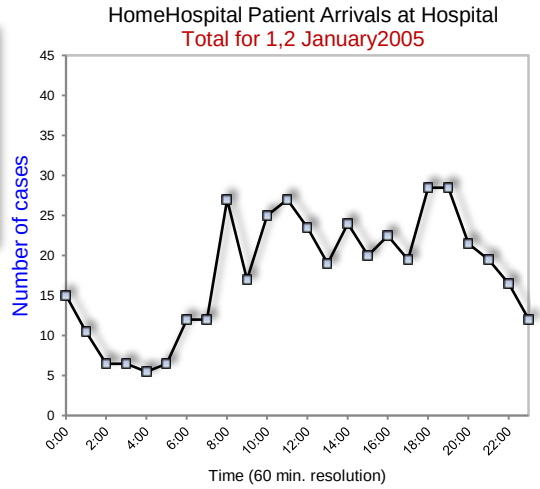
Excel 2003: click on tab Scale and fill in parameters.
Excel 2010: select Axis Options and fill in parameters.

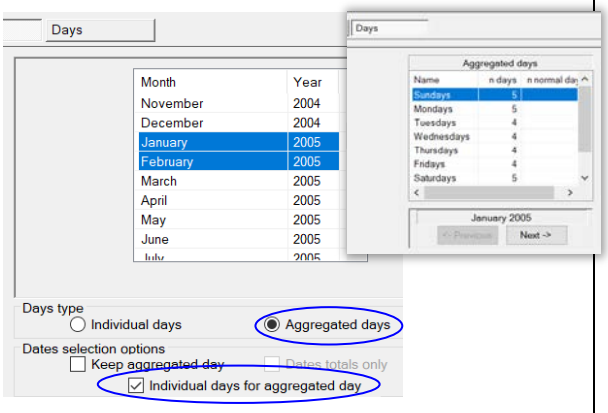
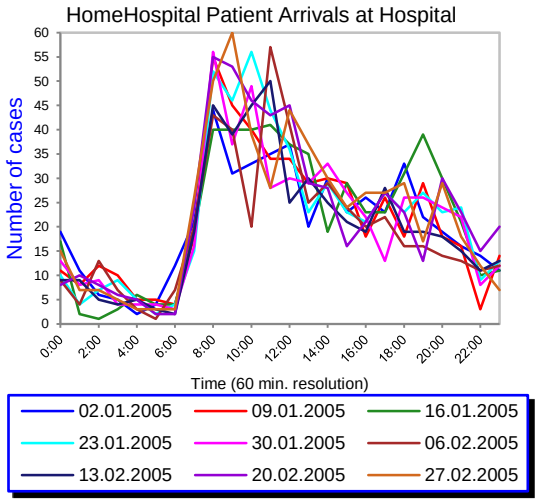
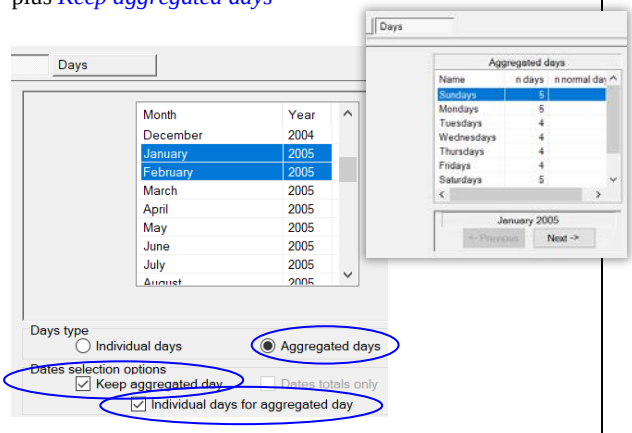
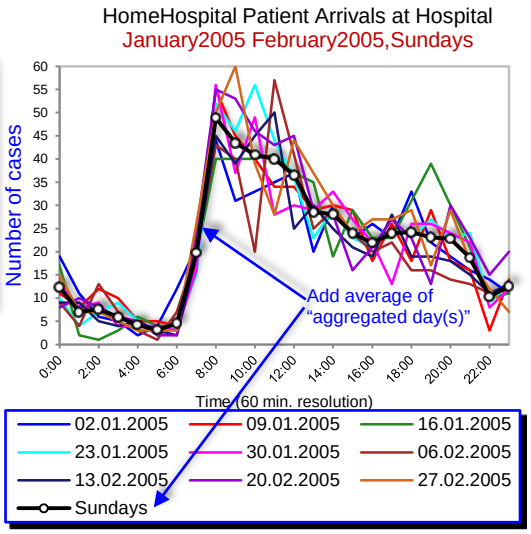
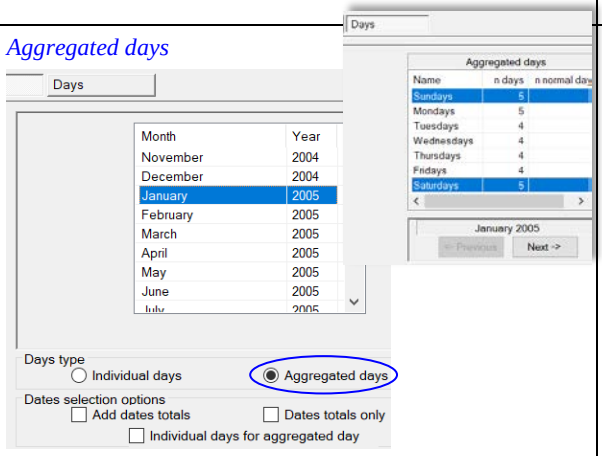
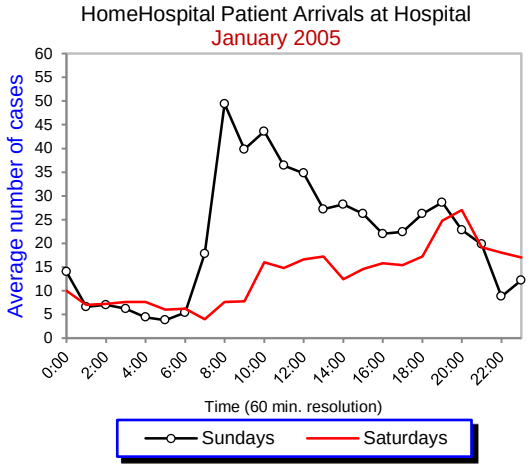


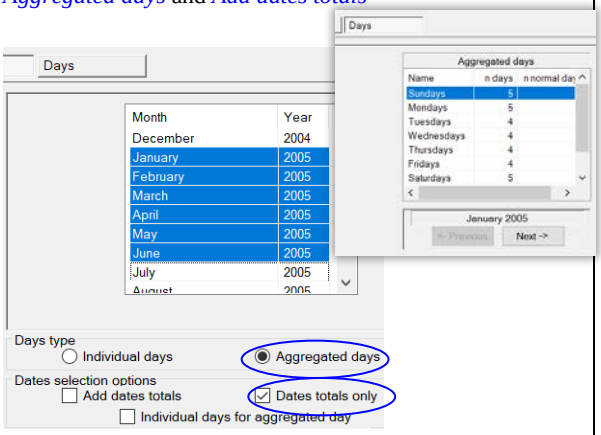
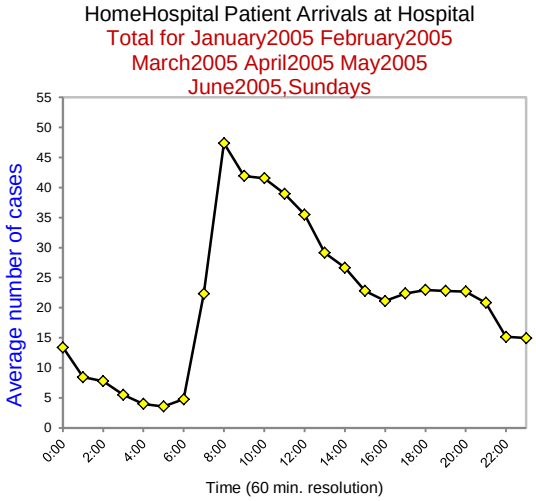
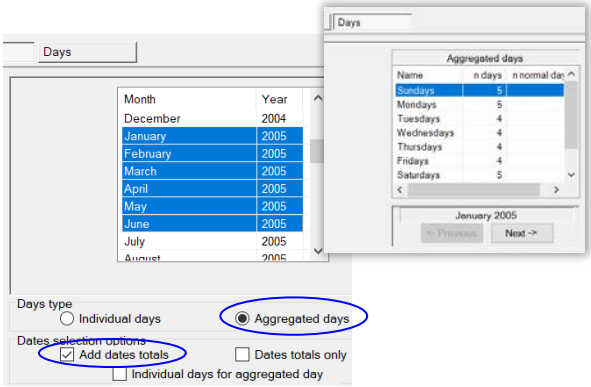
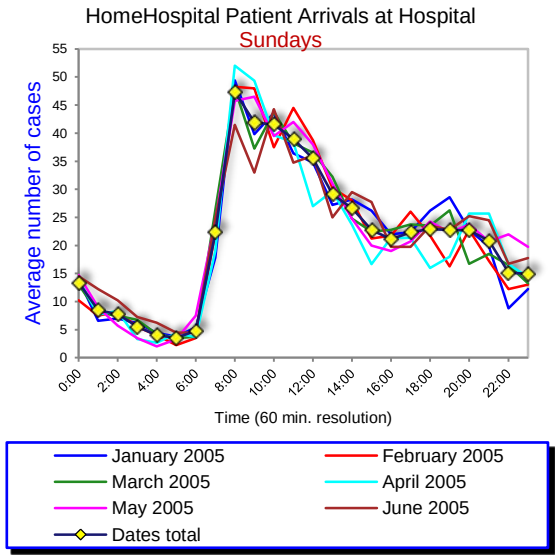
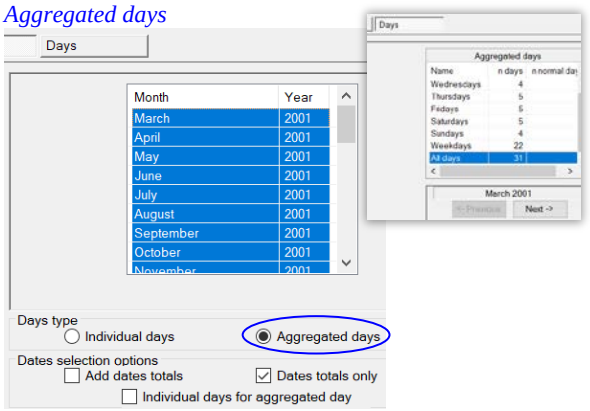
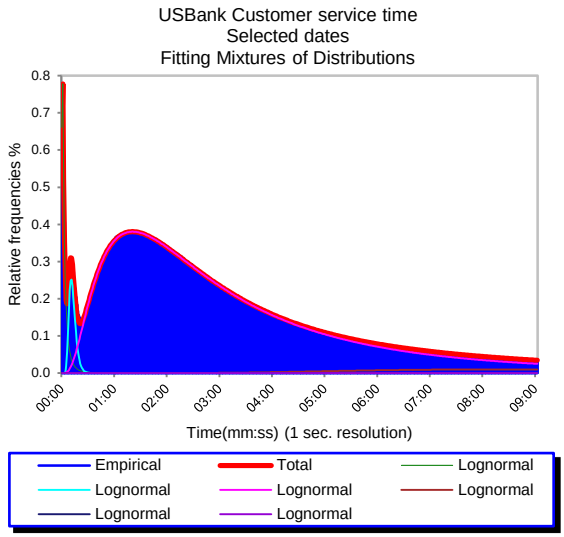
Excel 2016:


Format Axis
Axis Options
Text Options
Axis Options
Bounds
Minimum
0.0
Reset
Maximum
2.3
Reset
Units
Major
0.2
Reset
Minor
0.04
Auto
Horizontal axis crosses
Automatic
Axis value
0.0
Maximum axis value
Display units
None

Appendix E: How to Design a Sample for a SEEStat EDA

Data sample type	SEEStat	
	User Interface	Output Chart
Daily data - selection of specific day(s) based on date-in-month. In example: 2 charts – • 2 days – Saturday, January 1, 2005 and Sunday, January 2, 2005	*Individual days* 	
Daily data - selection of specific day(s) based on date-in-month, plus the average of these selected days. In example: 3 charts – • 2 days – Saturday, January 1, 2005 and Sunday, January 2, 2005 • Average (per hour) of these 2 selected days	*Individual days* and *Add dates totals* 	
Daily data - average of selected days. In example: 1 chart – • Average over the 2 selected days (Saturday, January 1, 2005 and Sunday, January 2, 2005)	*Individual days* and *Dates totals only* 	

Data sample type	SEEStat	
	User Interface	Output Chart
Daily data - selection of specific day(s)-of-week. In example: 9 charts – • 9 Sundays from January 2005 and February 2005	*Aggregated days* and *Individual days for aggregated day*  	
Daily data - selection of specific day(s)-of-week, as well as average of these selected days. In example: 10 charts – • 9 Sundays from January 2005 and February 2005 • Average (per hour) of these 9 Sundays	*Aggregated days* and *Individual days for aggregated day* plus *Keep aggregated days*  	
Daily average(s) – of selected specific day(s)-of-week, based on one or several months. In example: 2 charts – • Daily average based on Sundays from January 2005 • Daily average based on Saturdays from January 2005	*Aggregated days*  	

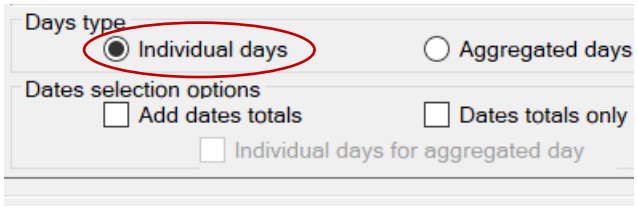
Data sample type	SEEStat	
	User Interface	Output Chart
Daily average(s) – of selected specific day(s)-of-week, based on one or several months. In example: 1 chart – • Daily average of Sundays during 6 months	*Aggregated days* and *Add dates totals*  	
Daily average(s) – of selected specific day(s)-of-week, based on one or several months; separately as well as average of all these selected months. In example: 7 charts – • 6 averages of Sundays, per each month separately • Average of all Sundays during 6 months	*Aggregated days* and *Add dates totals*  	
Select all data In example: 1 chart – • Based on all days during 32 months (the whole data) Statistics N: 41,988,545 N(average per day): 44,956 Mean: 4 min 5 sec Standard Deviation: 4 min 38 sec Median: 2 min 39 sec Skewness: 3.34 Kurtosis: 17.7 *Mixture fitting to about 42 million observations, of 6 LogNormal distributions, plus calculating summary statistics, takes approximately 14 seconds (on a reasonably strong laptop).*	*Aggregated days*  	

The above examples help you understand the functionality of terms used in your selections. We associate more rigorous interpretations with these terms:

"Days type" – there are two choice types for days:

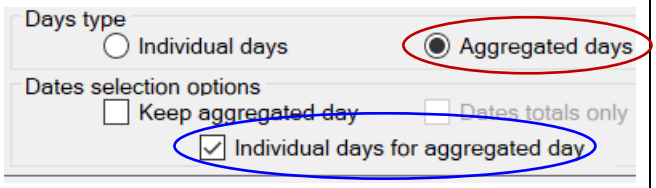
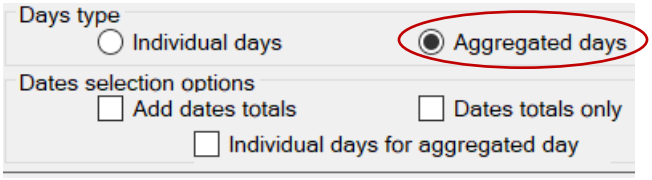
1. Individual days – chooses a specific day (e.g. Sep 11, 2001) or a set of selected days, based of date(s)-in-months.
2. Aggregated days – chooses specific day, or days, based on day-of-week (e.g. Sundays, or Wednesdays, or both)

1. When "Individual days" is selected, two "Dates selections options" appear:



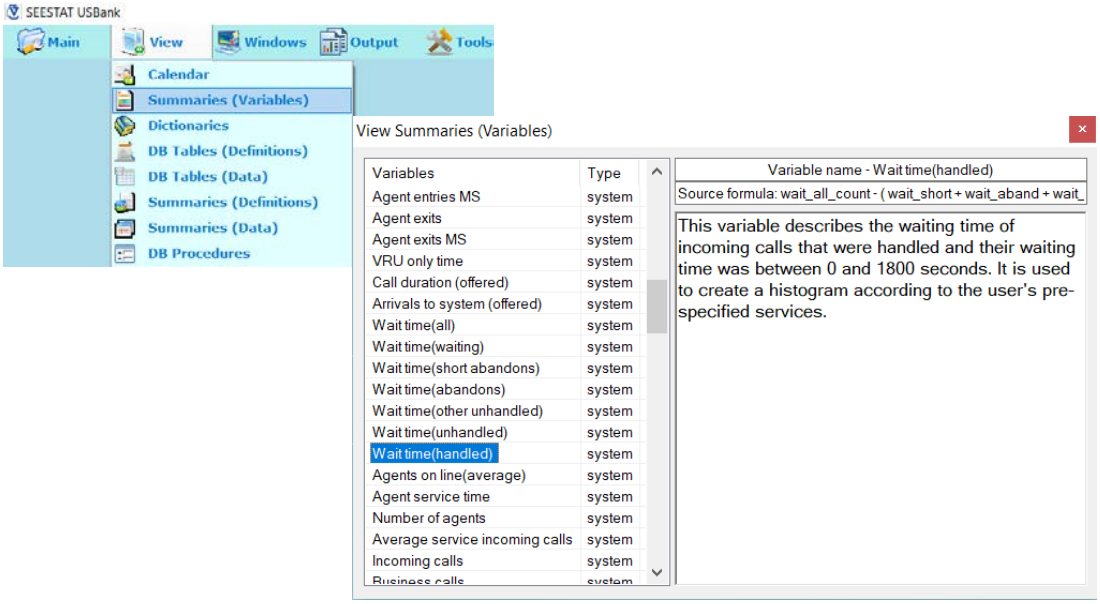
   a. Add dates totals – for specific days, chosen based on date-in-month, SEEStat plots individually day-in-month(s) and average of selected days in addition
   b. Dates totals only – displaying only average of the selected days

2. When "Aggregated days" is selected, there are two choices:

2.1 When "Individual days for aggregated day" is selected, one "Dates selections options" appears:	2.2 When Individual days for aggregated day is NOT selected, two "Dates selections options" appear:
	
Individual days for aggregated day – for chosen day(s)-of-week, and chosen month(s): SEEStat plots individually day-in-month(s) a. Keep aggregated day – supplements above by displaying average of selected days in addition	a. Add dates totals – for chosen day(s)-of-week, based on several months: SEEStat plots daily averages for each month separately, as well as average all these selected months b. Dates totals only – displaying average of all selected day(s)-of-week of selected months

Appendix F: Online Definitions of SEEStat Variables

Click **View**-> **Summaries (Variables)** and select variable.



Appendix G: The Offered-Load

A. *In the Summaries above, seek the description of the variable* ***Offered-Load****. It reads*:

The offered-load is a time-varying measure of the amount of work in a system. It accounts for both the arrival rate and the service times required by the arrival, including virtual services of abandoning customers. A way to animate the offered-load is as the least number of servers required to guarantee that no customer is delayed in queue upon arrival.

This (data-based) SEEStat-definition aims at covering both individual sample-paths as well as averages over multiple sample-paths. When modelling these two data-driven definitions, the first gives rise to a stochastic process, call it the *offered-load process* (or perhaps *the Workload* process), and the second corresponds to means of its paths – this is the prevailing model of the *offered-load*. Few more details about these models will be provided momentarily.

B. *Offered-load in SEEStat:*
The offered-load is a central measure of a service system. For example, it is the skeleton around which staffing plans are developed. One can calculate and display the offered-load in SEEStat, via a procedure that has some subtlety in it: note that the text above says "including virtual services of abandoning customers" – thus, calculating the offered-load entails estimating the would-have-been service-times of customers who in fact abandoned.
SEEStat does it the simple way: it simply samples, at random, from the collection of existing service times (of those customers who did not abandon).
But what if there is dependence between (im)patience and durations of service times? For example, it is plausible that customers are willing to wait less (more) for short (long) service times. Such dependence does indeed occur, which motivated a procedure estimating (would-have-been) service-durations given the time-willing-to-wait (which is observable). This is addressed in the following Technion MSc Thesis (advised jointly with Yaakov Ritov):
Reich M. **The Workload Process: Modelling, Inference and Applications**. M.Sc. Thesis, Technion, June 2012. (Thesis PDF-3MB) (Seminar PDF-2.4MB)

C. *Models of the Offered-Load – constant arrival rate:*
In the simplest setting of Poisson arrivals at a constant rate λ, iid service durations with mean E[S], and the usual assumptions of basic queueing models (e.g. independence between arrivals and services, and being in steady-state), the offered load is the familiar

$$\mathbf{R = \lambda \times E[S]} \quad (= \boldsymbol{\lambda/\mu}, \text{ in terms of service-rate } \boldsymbol{\mu} = 1/E[S]).$$

By Little's Law, R is also the average number of busy servers (served customers) in a naturally corresponding infinite-server model (Poisson arrivals with the above iid service durations): this infinite-server process is the offered-load process. Its mean, in steady-state, is the above R.

D. *Models of the Offered-Load – time-varying arrival rates:*
Significantly, the above definitions via infinite-server models carry over for time-varying arrival rates (e.g. time-inhomogeneous Poisson) – they are the ones that are commonly

used. It turns out that the offered-load enjoys beautiful (almost) explicit and insightful formulae. We shall briefly discuss them in class – indeed, they are a central subject in the references below.

Remark: In the spirit of "symmetry" between customers and servers, one can introduce "***Offered-Capacity***" to represent time-varying service capacities (again random per sample-path or averages of sample-paths). Specifically, given a time-varying *staffing-function*, the offered-capacity is the maximal number of served-customers so that no server remains idle upon service-completion (either sample-paths or average). It can be animated via a queueing station with an infinite-source of customers.

References

I recommend the following three papers by Ward Whitt and collaborators, in the presented order:

Relatively short introductory overview of staffing in a time-varying system:
"What You Should Know About Queueing Models To Set Staffing Requirements in Service Systems," 2007.
http://www.columbia.edu/~ww2040/shorter041907.pdf
Offered-load formulae: Section 6 in general and (6.1) in particular.

Some mathematical foundations - a beautiful paper about the time-varying infinite-server queue:
"The Physics of the M_t/G/infinity Queue," 1993.
http://www.columbia.edu/~ww2040/physics.pdf
e.g. Theorem 1 is the source of the formulae for the time-varying offered-load.

A recent 86-page survey, worthy of efforts by those seeking to study time-varying queues:
"Time-Varying Queues," 2018.
http://www.columbia.edu/~ww2040/TVQ_QMSM_062618.pdf