Empirical Analysis of a Call Center *

Avishai Mandelbaum, Anat Sakov and Sergey Zeltyn
Technion – Israel Institute of Technology †

March 25, 2001

Acknowledgment

We would like to thank Shlomo Philoshoph, the call-center manager of bank-anonymous, for first joining us in the belief that this project is feasible, and then helping us, always cheerfully and professionally, in making it happen.

We also thank Michael Yakobson and Elyashiv Kelerman, manager and programmer from the Composit company. Together with Ilan Guedj, they orchestrated the data transfer to the Technion. Ilan, now a PhD student at MIT, also designed the Internet site for the data, transformed the data from the format of Composit to the format in Table 1, and carried out the initial data debugging and cleaning.

Thanks are in order also to Ruth Grossman, now a PhD student at Tel-Aviv University. While working at the Statistical Laboratory at the Technion, Ruth fitted HEFT to preliminary versions of the data.

Larry Brown, Yaakov Ritov and Linda Zhao checked some of our statistical analysis, mostly confirming our findings but sometimes also refuting it, convincingly. With their help, we are hoping to have reached the right "picture".

Last but not least, useful conversations with Noah Gans, Haipeng Shen, and Nahum Shimkin are greatly appreciated.

*Supported by the ISF (Israeli Science Foundation) grant 388/99-2 (jointly with Nahum Shimkin, Technion EE), by Wharton's Financial Institutions Center, and by the Technion funds for the promotion of research and sponsored research.

†Faculty of Industrial Engineering and Management, Technion, Haifa, 32000, Israel; Emails: avim@tx.technion.ac.il, sakov@ie.technion.ac.il, zeltyn@ie.technion.ac.il.

Contents

1 Introduction

Call center is the common term for describing a telephone-based human-service operation. A call center provides *tele-services*, namely services in which the customers and the service agents are remote from each other. The agents, who sit in cubicles, constitute the physical embodiment of the call center: with numbers varying from very few to many hundreds, they serve customers over the phone, while facing a computer terminal that outputs and inputs customer data. The customers, who are only virtually present, are either being served, or they are waiting in, what we call, *tele-queues*: up to possibly thousands of customers sharing a phantom queue, invisible to each other and the agents serving them, waiting and accumulating impatience until one of two things happens – an agent is allocated to serve them (through a supporting software), or they *abandon* the tele-queue, plausibly due to impatience that has built up to exceed their anticipated worth of the service. The world of call centers is vast: some estimate [31] that 70% of all customer-business interactions occur in call centers; that $700 billions in goods and services were sold through call centers in 1997, and this figure has been expanding 20% annually; and that 3% of the U.S. working population is currently employed in call centers. (This amounts to 1.55 million agents, and some estimates actually go up to 6 million [5]). The leading-edge call center is a complex socio-technical system: its hunderds of agents could cater to thousands of customers per hour, in a way that average wait is measured in few seconds and agents' utilization exceeds 90%. Such simultaneous attainment of superb service quality with extreme resource efficiency is achieveable, despite ample stochastic variability, through scale-economies of unparalleled mangnitudes; and all this is possible only in the unique frictionless environment of computer-telephony integration and automatic call distribution.

Some view call centers as the *business frontiers* and others as the *sweat-shops* of the 21-st century. Either way, call centers provide ample uncharted challenges for researchers in multi-disciplines, from the soft (eg. Psychology, Sociology), through functional management (eg. Marketing, Information Systems), to the exact (eg. Computer Science, Mathematics). One should note that the challenges are, in fact, expanding: there exist an increasing number of multi-media call centers that can provide, in addition to the telephone, also video, Internet, fax and e.mail services. (The term customer *contact center* has been used to accommodate this broader connotation of a tele-service).

Our paper targets primarily researchers in Statistics, Operations-Research (especially Queueing Theory, and even more so Queueing Science – see Subsection 1.4.2), Operations Management, and Industrial Engineering. We believe that it is also of interest to researchers in telecommunications, and of use to managers that either run or oversee the operations of medium to large call centers.

1.1 Background and contents

The Call Center Magazine [26] is a U.S. monthly magazine (there are several others, for example Call Center Europe) that is dedicated to telephone services. Its readers are typically professionals in the call center industry. They are asked by the magazine to classify themselves according to the following business categories, which amply demonstrate the scope of telephone-services: advertising, banking, catalog retailer, computing, electronics or software, consulting, credit collection, direct mail marketer, dealer or distributor, entertainment, finance, securities or mutual funds, fund-raising, government, health-care, hospitality, information services, insurance, database supplier, manufacturer, market research, professional services, publishing or broadcasting, retailing, telecommunications, telemarketing, transportation, travel or recreation, utility, wholesaler, and more.

Sound scientific principles must be prerequisites for sustaining the levels of service and efficiency of today's call center, and these principles, in turn, better be based on real-world data. For example, in order to determine the least number of agents that could provide a given service level, it is critical to understand customers' (im)patience while waiting at the phone to be served (Garnet et.al. [12]). But such patience, as depticted in our Figures 12 and 13 for example, has *never* been documented. (Closely related exceptions are Palm [28] and Roberts [29] which could, unfortunately, be outdated). This collective ignorance typically leads to over-staffing (geared to "serve" those who abandon, over-simply put), which has severe economic consequences. To wit, the annual salary of a single agent is measured in several, often many, tens of thousands of dollars, and agents' salaries constitute about 70% of the cost of running a call center.

It is thus surprising, perhaps astonishing when considering the data-intensive hi-tech environment of the modern call centers, that operational data at the approriate resolution for research and management (as in Table 1), has been scarcely available. This is manifested by the lack of documented, comprehesive, empirical research of call centers, which is precisely our prime goal in this present study. Specifically, we have analyzed operational data of a bank call center, spanning all twelve months of 1999, at the level of individual telephone calls - this is a first of its kind, to the best of our knowledge. In addition, we seek to provide empirical foundations for Queueing Science, as in Zohar et.al. [35]. We also hope that our research constitutes a prototype that paves the way for future larger-scale studies, either at the individual call-center level, or perhaps even industry-wide. (At Wharton's Financial Institutions Center, such a study is now being initialized.) Finally, as data-quality is never perfect (see Table 9 and Section 10, for example), the present document should serve as a debugging and unifying device, for those wishing to analyze our or similar data.

Our data-base consists of over 440,000 individual telephone calls, each captured by a record that archives its event-history through the call center (see Table 1 and Figure 1). In contrast, prevalent call center data is only *averaged* over periods of fixed durations (15 minutes to a full day – See Subsection 1.2.1), which is not detailed enough for understanding call center characteristics. An example is customers' patience, the misunderstanding of which has significant negative consequences, as described above. In fact, the trigger for the present study happened to be practical consulting needs, which stimulated scientific curiosity for understanding human *tele-patience* at the phone.

One of us has been attempting to get hold of call center data, at the individual call-transaction level, for over four years (needless to say unsuccesfully till now). The present outcome, we hope and trust, provides a testimony that persistency has proved worthwhile, and that such data *can* indeed be archived, retrieved and analyzed. The knowledge gained, beyond being intellectually fascinating, must, in our opinion, accompany any scientifically-based engineering or management of call centers.

The data is described in Section 2 (see Table 1), and its analysis is carried out in Sections 3–9. We start with basic counts, then proceed with the analysis of arrivals (see, for example, Figures 2–6), waiting (eg. Figure 11), service durations (Figure 17) and patience. As already mentioned, the latter is of particular interest (eg. the hazard rates in Figure 12), being the first attempt at systematically recording and understanding tele-patience. Sections 8 and 9 clearly demonstrate the benefits from individual call data: the former is devoted to the behavior of individual customers and agents, and the latter to performance analysis of typical (and some atypical) days. Section 10 records some problematic records in the data. We conclude in Section 11 with directions for future research.

The rest of the Introduction is devoted first to a general discussion of call center data – its sources and suprising scarcity, then a desription of the particular bank call-center that has been the source for our data. We conclude with an appendix-like subsection on Service Engineering and Queueing Science (Subsection 1.4).

1.2 Call center data

We distinguish between three types of call center data: operational, marketing, and psychological. *Operational* data is typically collected by the Automatic Call Distributor (ACD), which is part of the telephony-switch infrastucture (typically hardware-, but recently more and more software-based). *Marketing* or *Business* data is gathered by the Computer Telephony Integration/Information (CTI) software, that connects the telephony-switch with company data-bases, typically customer profiles and business histories. Finally, *psychological* data is deduced from surveys of customers, agents or managers. It records subjective perceptions of service level and working environment, and will not be discussed here further. It is important to note that subjective survey data has also been used as a source for operational and marketing data. While serving a useful rough benchmarking role (see, for example, [2] and [21]), such data should be handled with care. It definitely can *not* serve as a substitute, or even a proxy, for the ACD and CTI data discussed below.

In this research, we analyze operational ACD data. The ultimate goal is, however, to integrate data from the three sources mentioned above, which is essential if one is to understand and quantify the role of (operational) service-quality as a driver for business success. However, there is ways to go. First, "dialogues" between ACDs and CTI's are non-existing (the two typically originate in separate vendors). Moreover, our experience has been that both types of data are very difficult to access: ACD data for technical reasons and CTI data due to confidentiality concerns. (Interestingly, this state of affairs may be different with Internet services – See Section 11 for an elaboration).

1.2.1 Operational data (ACD)

Most modern call centers are equipped with an ACD: this is the switch that routes calls to agents, while tracing and capturing the history of each call as it flows through the call center. ACD data include each call's arrival-time, waiting-time in the tele-queue, service duration, as in Table 1. (A related software tool goes under the name of Customer Relations Management (CRM) - it also records individual service transactions, but more in terms of work-content and customer-value rather than operational characteristics).

ACD data is typically used through aggregated reports. These consist of counts and averages over 15/30/60 minutes periods at the lowest level, and daily/weekly/yearly periods at higher levels. In such reports one can find, among other things, the total number of calls served or abandoned during the given period, average waiting times, the agents' utilization levels, etc. The call-log files, with individual call histories, is commonly *erased* after being aggregated. The reasons, we believe, are at least the following. First is the desire to save storage space, which nowadays is economically-unfounded: a whole month worth of data, from a large call center, would fit into a single compact disc. But more importantly is the lack of understanding on what can be done with individual-transaction data (the increasing popularity of CRM helps here) and sometimes also the lack of capabilities for deciphering vast data-warehouses (here Data Mining appears to come to some rescue.)

1.2.2 Marketing data (CTI)

The other main type of data is marketing (business) data. It is typically collected by CTI software (middleware), that integrates telephone data, specifically the caller ID, with computer data-bases that include the caller's profile and business-history. (Some associate CRM tools with this kind of data, rather than with ACD data). Having made the integration, the CTI software pops up a relevant description of the customer on the agent's terminal screen. This description includes, for example, the history of the previous calls and, if relevant, dollar-figures for past sales and future tele-marketing targets. (It is less relevant, for example, in Help Desks, which are Technical Support Centers; here history would inlclude, for example, past complaints and repairs).

1.3 The call center of Bank Anonymous

The source of our data is a small call center of one of Israel's banks. (The small size has proved convenient, in many respects, for a pioneering field study). The center provides several types of services: information for current and prospective customers, transactions of checking and saving accounts, stock-trading, and technical support for Internet users of the bank's site.

The call center consists of 8 regular-agent positions, 5 Internet-agent positions, and one shift-supervisor. Working hours are weekdays (Sunday to Thursday) from 7am to midnight; the center closes at 2pm on Friday and reopens around 8pm on Saturday.

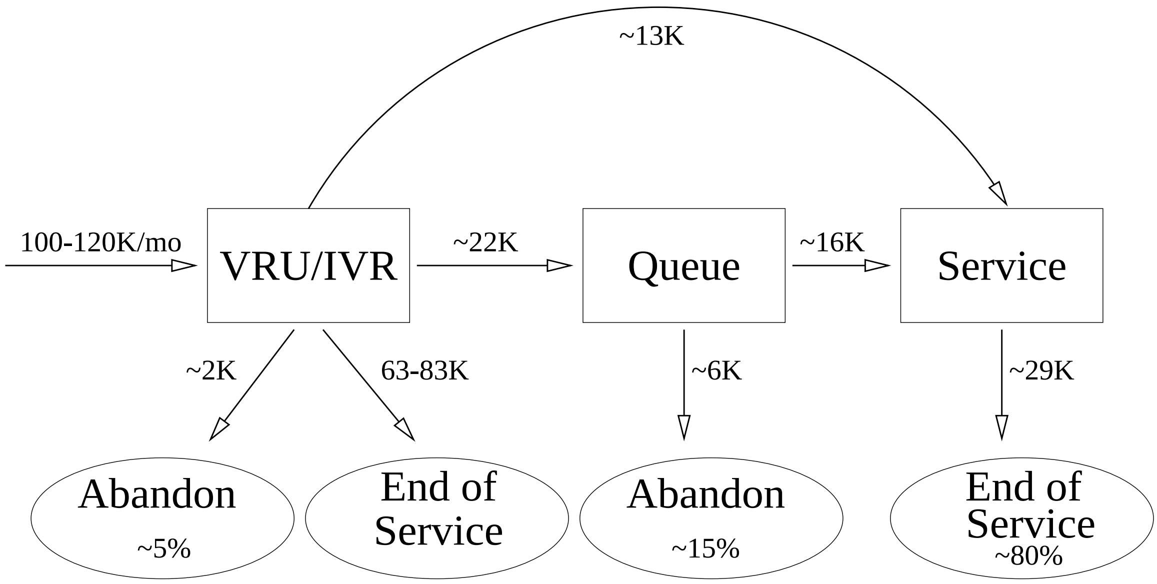
A simplified description of a call history is as follows. A customer calls one of several phone numbers of the call center, depending on the type of service sought. Except for rare busy-signals, the customer then "enters" the Voice Response Unit (VRU) (sometime called also Interactive Voice Response (IVR)). During this phase of the call, one must identify oneself, and then one gets some recorded information, general and customized (e.g. account balance). It is possible also to perform some self-service transactions at this stage and then complete one's service, as 65% of the customers actually do. The other 35% dialed originally a number that indicated their desire to speak to an agent. There are three options at this point: if there is a free agent who is capable of performing the desired service, the customer and the agent are matched to start service immediately; some customers actually abandon at this stage; the third option is to join the tele-queue.

Regular customers join the "end" of the tele-queue, while high-priority customers are advanced in the queue by 1.5 minutes right upon arrival. Service is then rendered on a first-come-first-served (FCFS) basis. While waiting, the caller receives some rather vague information, specifically the location in the queue and the amount of time that the first-in-queue has been waiting. (Location-in-queue in the sense that if there are N agents working, then the first N queueing customers are considered first in queue). The annoucement is replayed every 60 seconds or so, with music, news or commercials interwined. Most customers wait until either an appropriately-skilled agent becomes free, or else they abandon the tele-queue. Waiting customers can also choose to return to the VRU for self-service; their location in the queue is being saved, and when their turn arrives, they are transfered from the VRU to a free agent. (Such transfers between the queue and the VRU are not shown in the data).

There are 100,000–120,000 calls per month in total. Out of these, 30,000-40,000 seek to speak to an agent, and the remaining are satisfied with self-service transactions at the VRU. (The latter service operates 24 hours a day, 7 days a week). Our data consist of the calls who desired to speak to an agent. The data are compiled on a monthly basis, from January 1999 to December 1999.

In a schematic way, the event-history (process-flow) of a call is summarized in Figure 1. The numbers above or near the arrows represent the approximate number of calls, and the percents within the ovals represent the percent out of the calls in our data, at that stage. Specifically, the 100% consist of callers seeking service by an agent; out of these, 5% actually abandoned at the VRU, 15% abandoned the tele-queue and the rest were served. One distinguishes between *incoming* calls (a customer calling the system) and *outgoing* calls (the center calls a customer, which is regularly done for high-priority abandoning customers). The diagram corresponds to incoming calls. The only difference with respect to an outgoing call is that the latter has no arrow leading into the VRU.

Figure 1: Event history of an incoming call (units of rates are calls per month)



1.4 On Service Engineering, Queueing Science and call centers

We end the introductory part with a brief discussion of Service Engineering and Queueing Science, as we perceive them, in the context of call centers. The goal is to build up gradually to the point where we can justify our view of call centers as queueing systems. This view is admittedly biased by our “scientific origins” (Operations Research, Statistics), but we believe that it also has significant origin-independent merits.

1.4.1 Queues

Queues in services are often the arena where customers, service-providers (servers, or agents) and managers establish contact, in order to jointly create the service experience. Process-wise, queues play in services much the same role as inventories in manufacturing [7]. But in addition, "human queues" express preferences, complain, abandon and even spread around negative impressions. Thus, *customers* treat the queueing-experience as a window to the service-providing party, through which their judgement of it is shaped for better or worse. *Managers* can use queues as indicators (queues are the means, not the goals) for control and improvement opportunities. Indeed, queues provide unbiased quantifiable measures (these are not abundant in services), in terms of which performance is relatively easy to monitor and goals are naturally formulated. In summary, the measurement and modelling of service queues, and their design and management, is suggested here as an undertaking of both theoretical and practical significance.

1.4.2 Service Engineering and Queueing Science

We view the present study as part of broader research agenda, which we have been calling *Service Engineering*. The goal of Service Engineering is to develop scientifically-based design principles and tools (often culminating in software), that support and balance service quality and efficiency, from the likely conflicting perspectives of customers, servers, managers, and often also society. Queueing models constitute a natural convenient nurturing ground for the development of such principles and tools (eg. [12] and [4]). However, the existing supporting (Queueing) theory has been somewhat lacking, as will now be explained.

The bulk of what is called Queueing Theory, consists of research papers that formulate and analyze queueing models with realistic flavor. Most papers are knowledge-driven, where "solutions in search of a problem" are developed. Other papers are problem-driven, but most do not go far enough to a practical solution. Only some articles develop theory that is either rooted in or actually settles a real-world problem, and scarcely few carry the work as far as validating the model or the solution [13, 17]. In concert with this state of affairs, not much is available of what could be called Queueing Science, or perhaps the Science of Congestion, which should supplement traditional Queueing Theory with empirically-based models, observations and experiments. Two text books that acknowledge this view are Lee [25] and Hall [15]. In service networks, such "Science" is lagging behind that in telecommunications (Bertsekas and Gallager [3]), computers (Kleinrock [23]), transportation (Herman [19]) and manufacturing (Hopp and Spearman [20]). Key reasons for the gap seem to be the difficulty to measure service operations, combined with the need to incorporate human factors (which are notoriously difficult to quantify). We are hoping to take here some first steps towards closing this gap.

1.4.3 Call centers as queueing systems

Call centers can be beneficially viewed as stochastic systems, within the Operations Research paradigm of Queueing models. (Figure 1 has been inspired by this point of view). Queueing theory was conceived by Erlang at the beginning of the century [9, 10], and has flourished since to become one of the most central research themes of Operations Research [34, 14, 7]. In a queueing model of a call center, the customers are callers, servers (resources) are telephone agents (operators) or communication equipment, and queues consist of callers that await service by a system resource. The simplest and most-widely used such model is the M/M/S (Erlang-C) model. We refer the reader to [12, 17, 32, 11, 35] for some of its shortcomings. Granted these, for example the realization that service times need not be exponential and that abandonment is important to acknowledge, the questions that arise are about, for example, the statistical nature of service times and customer patience. These questions can and should be answered via research efforts such as the one reported here.

2 Data description

A sample of the data is given in Table 1. [1]

There is a record (a line in the file) for each call. The following are the fields for each record. We have changed the titles for some of the fields, and shortened some of the fields content to be able and fit the table on a single page (the changes will be described below). Fields 1 and 2 are call identifiers: the combination of 1 and 2 creates a unique identification for each call. Fields 3–5 provide information on the customer and fields 6–17 cover the call process flow.

1. Ln – Each entering phone-call is first routed through a VRU. There are 6 VRUs labeled AA01 to AA06. Each VRU has several lines labeled 1-16. There is a total of 65 lines. Each call is assigned a VRU number and a line number. Since all records in our sample data started with AA01, we omitted that part of the field.

2. Call – Each entering call is assigned a call ID. The ID's are not necessarily consecutive due to being assigned to different VRUs. All our call ID start with 44, so we omitted this from the ID.

3. Cust ID – identification number of the customer (caller). The ID is 0, if the caller is not identified by the system, as in the case for prospective customers. Due to a system bug, the ID was not recorded for customers who did not wait in queue (i.e. abandon from VRU, or reached an agent directly from VRU).

4. Pr – the priority of the customer is taken from an off-line file. The priority is 0 and 1 for unidentified and regular customers, and is 2 for priority customers. Customers are served in the order of their time in queue; however, priority customers are advanced in the queue by a minute and a half. Customers have not been told about the existence of priorities. Customers who did not wait in queue and whose ID is recorded as 0, have the value 0 recorded as their priority (even if this is not their true priority).

5. Tp – the type of service requested by the caller: regular activity (coded 'PS'), regular activity in English (coded 'PE'), Internet consulting (coded 'IN'), stock market (coded 'NE'), potential customers getting information (coded 'NW'). These are all inbound calls. Customers may leave their phone numbers in order to be called back. High priority customers who abandon are automatically called back. These are outbound calls, denoted by type TT. The dialing of TT calls is peformed by either agents or the computer. The call center has several phone numbers, each is associated with a different type of service. The system records the type of service according to the number dialed. When a customer calls the 'PS' number, the activity is recorded as 'PS', even if some other activity if performed, for example stock market trading. While there is no way for us to know the exact service contents, we were told that calling for one type while carrying out another is rare.

6. D – date of call in format year-month-day. In the sample data, we omitted the year (99) and month (09) from all calls. A typical entry is actually 990901.

7. VRU in – Time that the phone-call enters the call-center. More specifically, this is the time the call enters the VRU. All times are recorded in an hh:mm:ss (hours, minutes, seconds) format.

8. VRU out – Time of exit from the VRU: either to the queue, or directly to be served, or to leave the system (abandonment).

9. V – time (in seconds) spent in the VRU, calculated as the difference of the last two fields.

10. Q in – Time of joining the queue (being put on "hold"). This entry is 00:00:00, for customers who have not reached the queue (abandoned from the VRU).

[1] The data is available for analysis in `http://ie.technion.ac.il/Academ/Course/096324` (menu `Homework`, entry `Call Center Data`), with permission by A.M. (`avim@tx.technion.ac.il`).

Ln	Call	Cust ID	Pr	Tp	D	VRU in	VRU out	V	Q in	Q out	Q	Outcome	SER in	SER out	S	SERV ID
01	749	27644400	2	PS	01	11:45:33	11:45:39	6	11:45:39	11:46:58	79	AGENT	11:46:57	11:51:00	243	DORIT
01	750	12887816	1	PS	05	14:49:00	14:49:06	6	14:49:06	14:53:00	234	AGENT	14:52:59	14:54:29	90	ROTH
01	967	58660291	2	PS	05	14:58:42	14:58:48	6	14:58:48	15:02:31	223	AGENT	15:02:31	15:04:10	99	ROTH
01	968	0	0	NW	05	15:10:17	15:10:26	9	15:10:26	15:13:19	173	HANG	00:00:00	00:00:00	0	NO SER
01	969	63193346	2	PS	05	15:22:07	15:22:13	6	15:22:13	15:23:21	68	AGENT	15:23:20	15:25:25	125	STEREN
01	970	0	0	NW	05	15:31:33	15:31:47	14	00:00:00	00:00:00	0	AGENT	15:31:45	15:34:16	151	STEREN
01	971	41630443	2	PS	05	15:37:29	15:37:34	5	15:37:34	15:38:20	46	AGENT	15:38:18	15:40:56	158	TOVA
01	972	64185333	2	PS	05	15:44:32	15:44:37	5	15:44:37	15:47:57	200	AGENT	15:47:56	15:49:02	66	TOVA
01	973	3.06E+08	1	PS	05	15:53:05	15:53:11	6	15:53:11	15:56:39	208	AGENT	15:56:38	15:56:47	9	MORIAH
01	974	74780917	2	NE	05	15:59:34	15:59:40	6	15:59:40	16:02:33	173	AGENT	16:02:33	16:26:04	1411	ELI
01	975	55920755	2	PS	05	16:07:46	16:07:51	5	16:07:51	16:08:01	10	HANG	00:00:00	00:00:00	0	NO SER
01	976	0	0	NW	05	16:11:38	16:11:48	10	16:11:48	16:11:50	2	HANG	00:00:00	00:00:00	0	NO SER
01	977	33689787	2	PS	05	16:14:27	16:14:33	6	16:14:33	16:14:54	21	HANG	00:00:00	00:00:00	0	NO SER
01	978	23817067	2	PS	05	16:19:11	16:19:17	6	16:19:17	16:19:39	22	AGENT	16:19:38	16:21:57	139	TOVA
01	764	0	0	PS	01	15:03:26	15:03:36	10	00:00:00	00:00:00	0	AGENT	15:03:35	15:06:36	181	ZOHARI
01	765	25219700	2	PS	01	15:14:46	15:14:51	5	15:14:51	15:15:10	19	AGENT	15:15:09	15:17:00	111	SHARON
01	766	0	0	PS	01	15:25:48	15:26:00	12	00:00:00	00:00:00	0	AGENT	15:25:59	15:28:15	136	ANAT
01	767	58859752	2	PS	01	15:34:57	15:35:03	6	15:35:03	15:35:14	11	AGENT	15:35:13	15:35:15	2	MORIAH
01	768	0	0	PS	01	15:46:30	15:46:39	9	00:00:00	00:00:00	0	AGENT	15:46:38	15:51:51	313	ANAT
01	769	78191137	2	PS	01	15:56:03	15:56:09	6	15:56:09	15:56:28	19	AGENT	15:56:28	15:59:02	154	MORIAH
01	770	0	0	PS	01	16:14:31	16:14:46	15	00:00:00	00:00:00	0	AGENT	16:14:44	16:16:02	78	BENSION
01	771	0	0	PS	01	16:38:59	16:39:12	13	00:00:00	00:00:00	0	AGENT	16:39:11	16:43:35	264	VICKY
01	772	0	0	PS	01	16:51:40	16:51:50	10	00:00:00	00:00:00	0	AGENT	16:51:49	16:53:52	123	ANAT
01	773	0	0	PS	01	17:02:19	17:02:28	9	00:00:00	00:00:00	0	AGENT	17:02:28	17:07:42	314	VICKY
01	774	32387482	1	PS	01	17:18:18	17:18:24	6	17:18:24	17:19:01	37	AGENT	17:19:00	17:19:35	35	VICKY
01	775	0	0	PS	01	17:38:53	17:39:05	12	00:00:00	00:00:00	0	AGENT	17:39:04	17:40:43	99	TOVA
01	776	0	0	PS	01	17:52:59	17:53:09	10	00:00:00	00:00:00	0	AGENT	17:53:08	17:53:09	1	NO SER
01	777	37635950	2	PS	01	18:15:47	18:15:52	5	18:15:52	18:16:57	65	AGENT	18:16:56	18:18:48	112	ANAT
01	778	0	0	NE	01	18:30:43	18:30:52	9	00:00:00	00:00:00	0	AGENT	18:30:51	18:30:54	3	MORIAH
01	779	0	0	PS	01	18:51:47	18:52:02	15	00:00:00	00:00:00	0	AGENT	18:52:02	18:55:30	208	TOVA
01	780	0	0	PS	01	19:19:04	19:19:17	13	00:00:00	00:00:00	0	AGENT	19:19:15	19:20:20	65	MEIR
01	781	0	0	PS	01	19:39:19	19:39:30	11	00:00:00	00:00:00	0	AGENT	19:39:29	19:41:42	133	BENSION
01	782	0	0	NW	01	20:08:13	20:08:25	12	00:00:00	00:00:00	0	AGENT	20:08:28	20:08:41	13	NO SER
01	783	0	0	PS	01	20:23:51	20:24:05	14	00:00:00	00:00:00	0	AGENT	20:24:04	20:24:33	29	BENSION
01	784	0	0	NW	01	20:36:54	20:37:14	20	00:00:00	00:00:00	0	AGENT	20:37:13	20:38:07	54	BENSION
01	785	0	0	PS	01	20:50:07	20:50:16	9	00:00:00	00:00:00	0	AGENT	20:50:15	20:51:32	77	BENSION
01	786	0	0	PS	01	21:04:41	21:04:51	10	00:00:00	00:00:00	0	AGENT	21:04:50	21:05:59	69	TOVA
01	787	0	0	PS	01	21:25:00	21:25:13	13	00:00:00	00:00:00	0	AGENT	21:25:13	21:28:03	170	AVI
01	788	0	0	PS	01	21:50:40	21:50:54	14	00:00:00	00:00:00	0	AGENT	21:50:54	21:51:55	61	AVI
01	789	9103060	2	NE	01	22:05:40	22:05:46	6	22:05:46	22:09:52	246	AGENT	22:09:51	22:13:41	230	AVI
01	790	14558621	2	PS	01	22:24:11	22:24:17	6	22:24:17	22:26:16	119	AGENT	22:26:15	22:27:28	73	VICKY

Table 1: Sample of Data

11. Q out – Time of exiting the queue: either to receive service or to abandon.
12. Q – Time spent in queue, calculated as the difference between the last two fields.
13. Outcome – there are three possible outcomes: AGENT when a customer receives service; HANG when a customer hangs up. The third possible outcome PHANTOM is, according to the call center staff, abandonment from the VRU. However, the data does not support this, as more than 95% of the PHANTOM calls have positive service time. Hence, when considering the time a customer is willing to wait before abandoning, the virtual waiting time (see Section 6.3), and the analysis of individual days (Section 9), we ignore records with PHANTOM outcome.
14. SER in – Time of beginning of service by agent.
15. SER out – Time of end of service by agent.
16. S – Service duration, calculated as difference of last two fields.
17. Serv ID – the name of the agent who served the caller. If no service was provided this field is NO SER (In the data we have NO SERVER).

3 Basic counts

We start with some basic counts of calls, segmented by different covariates. Then, in Section 4 we continue with an analysis of the arrival process, displaying it in varying resolutions and segmenting it similarly.

Calls are segmented, for example, by outcome (HANG, AGENT, PHANTOM), service type (PS, NW, NE, IN, TT, PE), whether the caller is identified or not, etc. Some of our segmentations are interesting in their own right. Others have been found to be useful in debugging the data, especially those in Tables 5–9. For example, through these segmentations we identified data inconsistencies that, consequently, pointed to bugs in the program that assembled the bank data. This led, ultimately, to a replacement of 4 months worth of data.

In all tables, the numbers in parenthesis are the percent out of the column total. (-) represent a percentage smaller than 0.05.

Table 2: Monthly call counts (% out of yearly total)

Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Total
31599	33344	38801	32036	39553	37935	39022	42078	31371	34625	41019	43065	444448
(7.1)	(7.5)	(8.7)	(7.2)	(8.9)	(8.5)	(8.8)	(9.5)	(7.1)	(7.8)	(9.2)	(9.7)	(100)

Table 3: Call Counts by outcome (A – AGENT; H – HANG; P – PHANTOM)

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Total
A	27060	27162	27847	23583	29923	31193	29596	31402	27062	30861	33371	34083	353143
	(85.6)	(81.5)	(71.7)	(73.6)	(75.6)	(82.2)	(75.9)	(74.6)	(86.3)	(89.1)	(81.3)	(79.1)	(79.5)
H	4299	5904	10547	8148	9212	6420	9048	10366	4146	3617	7351	8648	87706
	(13.6)	(17.7)	(27.2)	(25.4)	(23.3)	(16.9)	(23.2)	(24.6)	(13.2)	(10.4)	(17.9)	(20.1)	(19.7)
P	240	278	407	305	418	322	378	310	163	147	297	334	3599
	(0.8)	(0.8)	(1)	(1)	(1.1)	(0.8)	(1)	(0.7)	(0.5)	(0.4)	(0.7)	(0.8)	(0.8)
Tot	31599	33344	38801	32036	39553	37935	39022	42078	31371	34625	41019	43065	444448

Table 4: Call counts by waiting status ($Q > 0$ or $Q = 0$)

Q	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Total
$= 0$	14694	15038	10635	10691	13031	15942	11975	16116	17704	22578	16643	16176	181223
	(46.5)	(45.1)	(27.4)	(33.4)	(32.9)	(42)	(30.7)	(38.3)	(56.4)	(65.2)	(40.6)	(37.6)	(40.8)
> 0	16905	18306	28166	21345	26522	21993	27047	25962	13667	12047	24376	26889	263225
	(53.5)	(54.9)	(72.6)	(66.6)	(67.1)	(58)	(69.3)	(61.7)	(43.6)	(34.8)	(59.4)	(62.4)	(59.2)
Tot	31599	33344	38801	32036	39553	37935	39022	42078	31371	34625	41019	43065	444448

Table 5: Call counts by being identified – ($ID \neq 0$) or not ($ID = 0$)

ID	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Total
$= 0$	18175	19558	15436	13827	17837	20047	17175	23030	21071	26082	21109	21205	234552
	(57.5)	(58.7)	(39.8)	(43.2)	(45.1)	(52.8)	(44)	(54.7)	(67.2)	(75.3)	(51.5)	(49.2)	(52.8)
$\neq 0$	13424	13786	23365	18209	21716	17888	21847	19048	10300	8543	19910	21860	209896
	(42.5)	(41.3)	(60.2)	(56.8)	(54.9)	(47.2)	(56)	(45.3)	(32.8)	(24.7)	(48.5)	(50.8)	(47.2)
Tot	31599	33344	38801	32036	39553	37935	39022	42078	31371	34625	41019	43065	444448

Table 6: Counts stratification by waiting status and identification

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Total
ID = 0,	14220	14520	9606	9916	12326	15452	11450	15826	17584	22494	16295	15721	175410
Q = 0	(45)	(43.5)	(24.8)	(31)	(31.2)	(40.7)	(29.3)	(37.6)	(56.1)	(65)	(39.7)	(36.5)	(39.5)
ID = 0,	3955	5038	5830	3911	5511	4595	5725	7204	3487	3588	4814	5484	59142
Q > 0	(12.5)	(15.1)	(15)	(12.2)	(13.9)	(12.1)	(14.7)	(17.1)	(11.1)	(10.4)	(11.7)	(12.7)	(13.3)
ID ≠ 0,	474	518	1029	775	705	490	525	290	120	84	348	455	5813
Q = 0	(1.5)	(1.6)	(2.7)	(2.4)	(1.8)	(1.3)	(1.3)	(0.7)	(0.4)	(0.2)	(0.8)	(1.1)	(1.3)
ID ≠ 0,	12950	13268	22336	17434	21011	17398	21322	18758	10180	8459	19562	21405	204083
Q > 0	(41)	(39.8)	(57.6)	(54.4)	(53.1)	(45.9)	(54.6)	(44.6)	(32.5)	(24.4)	(47.7)	(49.7)	(45.9)
Tot	31599	33344	38801	32036	39553	37935	39022	42078	31371	34625	41019	43065	444448

Table 3 splits the counts according to the outcomes as reported in the data. However, a customer can reach an agent directly from the VRU (zero waiting time) or after waiting in the queue. A customer may abandon from the queue (which should be associated with outcome HANG and positive waiting time) or from the VRU (outcome is HANG and zero waiting time). Table 7 summarizes the counts for the five possible feasible outcomes: AGENT with Q = 0, AGENT with Q > 0, HANG with with Q = 0, HANG with Q > 0 and PHANTOM.

Table 7: Refined outcome counts (A – AGENT; H – HANG; P – PHANTOM and Q = 0 or Q > 0)

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Total
A	13475	13066	8669	8202	10892	14157	9714	13010	16113	21050	14787	14001	157136
= 0	(42.6)	(39.2)	(22.3)	(25.6)	(27.5)	(37.3)	(24.9)	(30.9)	(51.4)	(60.8)	(36)	(32.5)	(35.4)
A	13585	14096	19178	15381	19031	17036	19882	18392	10949	9811	18584	20082	196007
> 0	(43)	(42.3)	(49.4)	(48)	(48.1)	(44.9)	(51)	(43.7)	(34.9)	(28.3)	(45.3)	(46.6)	(44.1)
H	1219	1972	1966	2489	2139	1785	2261	3106	1591	1528	1856	2175	24087
= 0	(3.9)	(5.9)	(5.1)	(7.8)	(5.4)	(4.7)	(5.8)	(7.4)	(5.1)	(4.4)	(4.5)	(5.1)	(5.4)
H	3080	3932	8581	5659	7073	4635	6787	7260	2555	2089	5495	6473	63619
> 0	(9.7)	(11.8)	(22.1)	(17.7)	(17.9)	(12.2)	(17.4)	(17.3)	(8.1)	(6)	(13.4)	(15)	(14.3)
P	240	278	407	305	418	322	378	310	163	147	297	334	3599
	(0.8)	(0.8)	(1)	(1)	(1.1)	(0.8)	(1)	(0.7)	(0.5)	(0.4)	(0.7)	(0.8)	(0.8)
Tot	31599	33344	38801	32036	39553	37935	39022	42078	31371	34625	41019	43065	444448

We now distinguish between incoming calls (types PS,NE,NW,IN and PE) and outgoing calls (TT). Table 8 is identical to Table 7 but restricted to type TT only.

Table 8: Refined outcome counts, for type TT (A – AGENT; H – HANG; P – PHANTOM and Q = 0 or Q > 0)

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Total
A	602	729	1398	1019	894	745	763	523	208	226	616	757	8480
= 0	(61.9)	(67)	(69.8)	(73.8)	(69)	(69)	(67.5)	(72.8)	(62.5)	(71.7)	(67.4)	(71.1)	(69)
A	250	255	383	234	252	244	238	117	88	63	190	194	2508
> 0	(25.7)	(23.4)	(19.1)	(17)	(19.4)	(22.6)	(21)	(16.3)	(26.4)	(20)	(20.8)	(18.2)	(20.4)
H	13	7	7	21	3	4	3	6	1	0	1	13	79
= 0	(1.3)	(0.6)	(0.3)	(1.5)	(0.2)	(0.4)	(0.3)	(0.8)	(0.3)	(0)	(0.1)	(1.2)	(0.6)
H	62	56	127	61	64	49	68	61	34	22	106	93	803
> 0	(6.4)	(5.1)	(6.3)	(4.4)	(4.9)	(4.5)	(6)	(8.5)	(10.2)	(7)	(11.6)	(8.7)	(6.5)
P	45	41	88	45	83	38	59	11	2	4	1	8	425
	(4.6)	(3.8)	(4.4)	(3.3)	(6.4)	(3.5)	(5.2)	(1.5)	(0.6)	(1.3)	(0.1)	(0.8)	(3.5)
Tot	972	1088	2003	1380	1296	1080	1131	718	333	315	914	1065	(12295)

Note that there are less abandonment for type ‘TT’, and that the fraction of calls reaching an agent without waiting in queue is about double that of the overall fraction. Still, it is important and instructive to reflect on the meaning of positive waits for TT customers. Many such calls are outbound calls, initiated by the shift supervisor, performed automatically by a dialing computer, and aimed at high-priority customers who abandoned: a customer involved in such a call, encountered a full queue (all agents are busy) and was forced to wait. Such an event happened 3,311 (too many) times during 1999, out of which 24.25% (803 customers) abandoned. Somewhat surprisingly, this latter figure is almost identical to the yearly fraction of abandonment, 24.17% out of delayed customers. (One would expect the TT fraction to be higher, because a dialed-to customer does not anticipate to be put on hold.)

Three possible sources contribute to the total time spent in the system (sojourn time): VRU time, queue time and service time. Table 9 has been proved most important for debugging our data-base.

Table 9: Combinations of VRU, queue and service times

VRU	< 0		$= 0$				> 0				
q	> 0		$= 0$		> 0		$= 0$		> 0		
SER	$= 0$	> 0	$= 0$	> 0	$= 0$	> 0	$= 0$	> 0	$= 0$	> 0	Total
Jan	11	22	328	2	127	189	1064	13300	2978	13578	31599
Feb	2	23	355	2	55	255	1653	13028	3828	14143	33344
Mar	8	19	608	0	107	419	1380	8647	8420	19193	38801
Apr	6	31	600	2	44	224	1902	8187	5561	15479	32036
May	30	48	558	2	48	268	1601	10870	6913	19215	39553
Jun	4	23	433	1	47	264	1396	14112	4563	17092	37935
Jul	12	34	598	2	55	264	1692	9683	6738	19944	39022
Aug	4	15	547	1	61	124	2602	12966	7151	18607	42078
Sep	4	10	279	0	29	81	1370	16055	2529	11014	31371
Oct	2	2	186	2	19	63	1370	21020	2039	9922	34625
Nov	2	18	408	1	100	174	1463	14771	5338	18744	41019
Dec	3	17	490	0	93	181	1708	13978	6338	20257	43065
Tot	88	262	5390	15	785	2506	19201	156617	62396	197188	444448
	(-)	(.1)	(1.2)	(-)	(0.2)	(0.6)	(4.3)	(35.2)	(14)	(44.4)	(100)

Most of the calls with zero VRU time and either positive time in queue or positive service time or both, are associated with type 'TT' calls: these are outgoing calls, initiated by the center so that the customer called-to need not go through the VRU. We are unclear on the source of negative VRU time, or zero VRU times for types other than 'TT' (which are, moreover, accompanied by positive service time or waiting time).

4 The arrival process

The arrival process records the epochs that phone calls arrive to the call center. It can be described at different levels of detail, and from various points of view. In this paper we provide only deterministic "fluid-like" descriptions of arrivals, which arise from averaging out stochastic variability. We leave for future research the statistical characterization of arrivals. (For example: does a time-inhomogeneous Poisson model fit the daily arrival process? if so, how accurate is the fit, and if not, what does?)

The first subsection provides a hierarchy of descriptions for arrivals, which differ in their resolution: yearly, monthly/weekly, daily and hourly. In the second subsection, arrivals are stratified according to customer types and prioirities. Daily desciptions of arrivals are further pursued in Section 9.

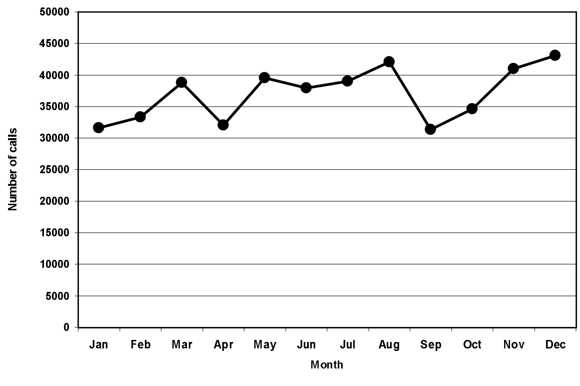
4.1 Hierarchical profiles: strategic, tactical, operational

The arrival process will now be described at four levels of representation, which differ by their time-scale as in Buffa, Cosgrove and Luce [6]. The three top levels also correspond to the classical hierarchical levels of decision making, proposed by Anthony [1]: Figure 2 is a top-level yearly picture, with month as the time unit, that supports strategic decisions; Figure 3 is a middle-level monthly picture with day as a unit, that supports tactical decisions; and Figure 4 is a daily picture, with unit hours, that supports operational decisions. In a typical call center, all three figures would exhibit *predictable* variability, in the sense that, for example, repeating Figure 3 for each month, as done in Figure 6, yields a predictable pattern. In contrast, Figure 5 is an hourly picture, with minutes as a time unit, that depicts *stochastic* or random variability. We shall provide momentarily a more detailed description of the figures, then continue with several segementations of the arrival process.

Hierarchical decision making is required, for example, to support the complex task of staffing a call center. At the top level, one must decide on how many agents are needed all in all, perhaps by season, which affects hiring and training. At a lower level one determines a shift structure over the month, which is determined in turn by daily and hourly staffing levels. Hourly staffing levels, or FTE's (full-time-equivalent) are commonly determined via queueing models that tradeoff service-quality against agents' efficiency. At their simplest form, staffing algorithms are described well already in [6]. The needs of the modern call center, however, go far beyond [6], in fact beyond state-of-the-art research, as described in [11].

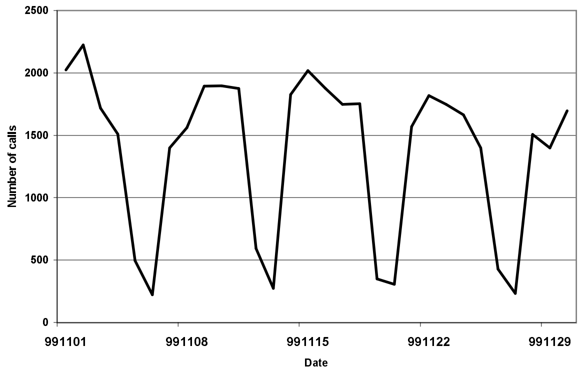
Figure 2 shows the number of calls per month, during 1999. Responding to changes in it at a specific call center would require strategic decisions. Note the decrease in number of calls in April and September, which is due to many holidays.

Figure 2: Strategical level. Number of calls per month



The next level displays the number of calls per day over a month, specifically November in Figure 3. The "valleys" occur during weekends, where the center operates only for a small number of hours. The picture for other months is similar (Figure 6), with additional valleys during holidays. (Examples are Yom-Kipor and Rosh-HaShana in September, and Passover and Independence Day in April. The first holiday in Passover was March 31st to April 1st, hence it is harder to notice on a monthly scale). This is a tactical-level figure: given the total number of agents available, their assignments must be made according to weekdays and weekends/holidays. To this end, it is also useful to add a tactical weekly picture.

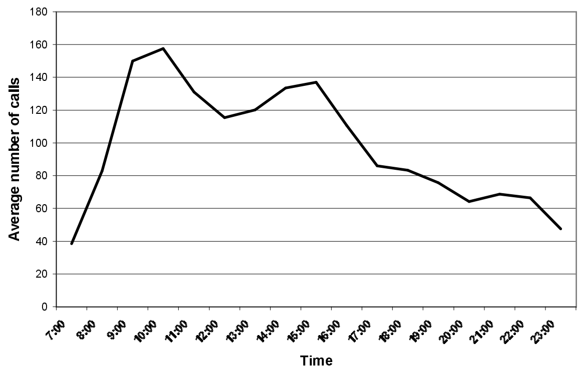
Figure 3: Tactical level. Number of calls per day (Nov)



At the operational level, staffing is made to fit peaks ("rush hours") and valleys. Figure 4 shows the average number of calls per hour during weekdays in November. Clearly the system is most busy around 9am, then the number of arriving calls decreases gradually till around noon, and increases again till about 4pm (closing time of the stock market).

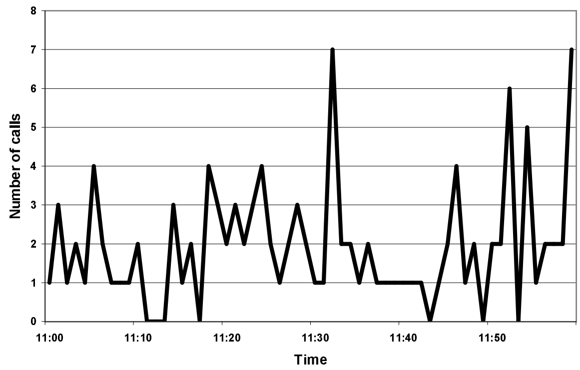
The operational level is of prime interest to a significant part of our targeted readers. It is therefore dwelt on in Section 9.2, where we focus on a typical day, followed by an analysis of two unusual days. It is interesting to compare Figure 4, displaying predictable variability over a day, with Figures 36 and 41, corresponding to unpredictable (the two unusual) days.

Figure 4: Operational level. Average number of calls per hour (Nov, weekdays)



Finally, when looking at an individual hour, calls seem to arrive randomly. Figure 5, which manifests this stochastic variability, displays the number of calls per minute, that arrive between 11am and 12pm during one typical day in November (November 7th). It is now clear that predictable variability emerges from stochastic variability by averaging the latter out.

Figure 5: Stochastic level. Number of calls per minute (Nov $7th$)



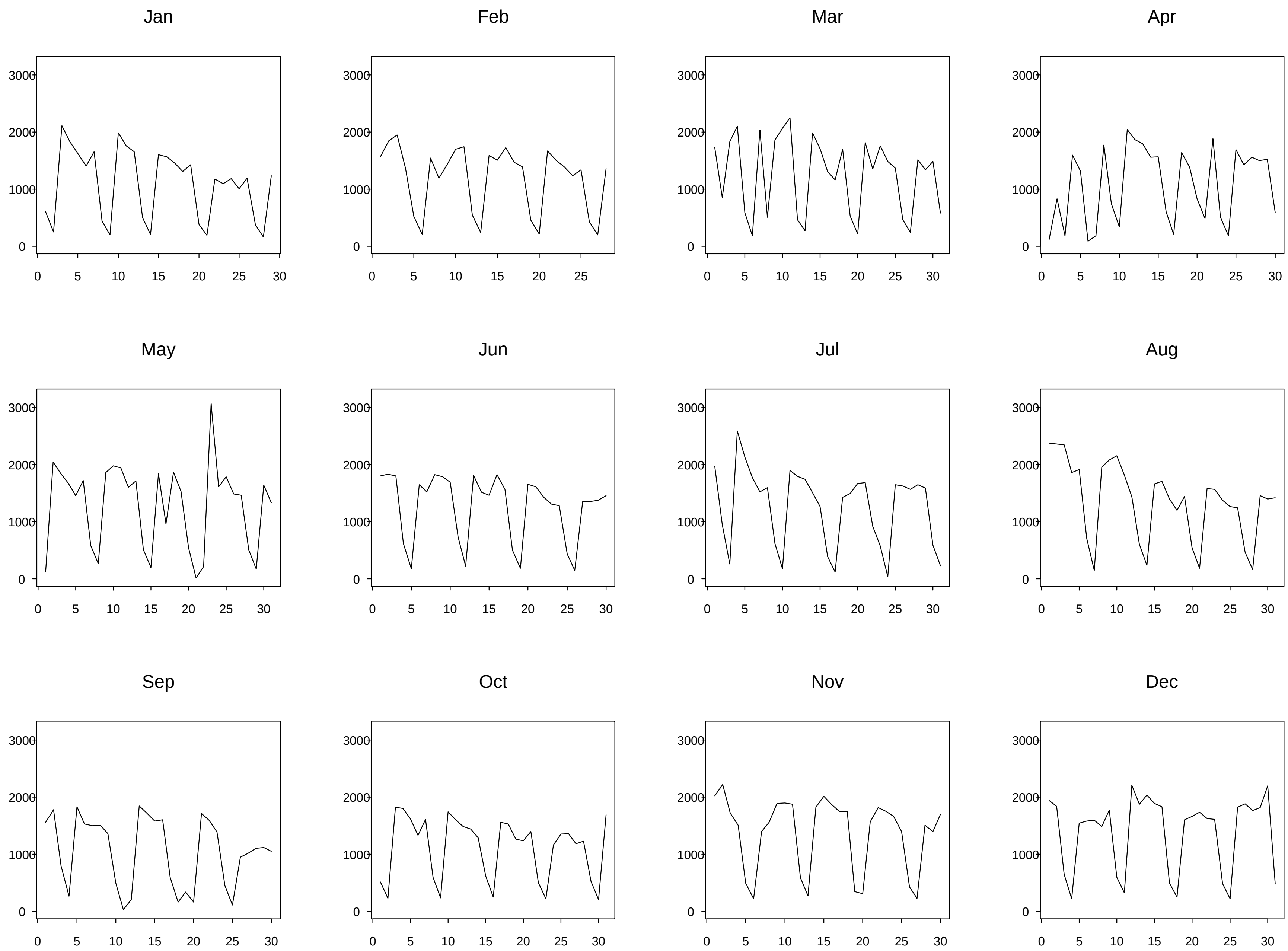


Figure 6: Number of calls per day

4.2 Customer profile counts: types and priorities

It is of interest to analyze the distribution of calls according to the different types of service and customer priorities. We do this in the present section.

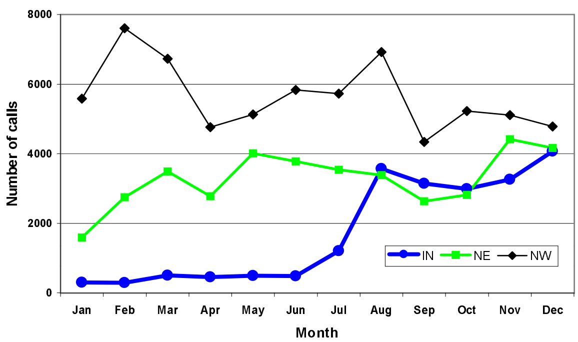
Table 10: Call counts by service type

Tp	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Total
PS	22998	21460	25910	22551	28482	26608	27269	27312	20810	23145	27156	28821	302522
	(72.8)	(64.4)	(66.8)	(70.4)	(72)	(70.1)	(69.9)	(64.9)	(66.3)	(66.8)	(66.2)	(66.9)	(68.1)
NW	5582	7604	6732	4760	5126	5832	5723	6923	4340	5222	5104	4780	67728
	(17.7)	(22.8)	(17.4)	(14.9)	(13)	(15.4)	(14.7)	(16.5)	(13.8)	(15.1)	(12.4)	(11.1)	(15.2)
NE	1586	2750	3490	2776	4011	3776	3536	3381	2635	2815	4418	4168	39342
	(5)	(8.2)	(9)	(8.7)	(10.1)	(10)	(9.1)	(8)	(8.4)	(8.1)	(10.8)	(9.7)	(8.9)
IN	295	293	503	451	492	483	1202	3563	3139	2987	3257	4067	20732
	(0.9)	(0.9)	(1.3)	(1.4)	(1.2)	(1.3)	(3.1)	(8.5)	(10)	(8.6)	(7.9)	(9.4)	(4.7)
TT	972	1088	2003	1380	1296	1080	1131	718	333	315	914	1065	12295
	(3.1)	(3.3)	(5.2)	(4.3)	(3.3)	(2.8)	(2.9)	(1.7)	(1.1)	(0.9)	(2.2)	(2.5)	(2.8)
PE	164	149	163	117	146	156	160	181	113	141	170	164	1824
	(0.5)	(0.4)	(0.4)	(0.4)	(0.4)	(0.4)	(0.4)	(0.4)	(0.4)	(0.4)	(0.4)	(0.4)	(0.4)
Tot	31597	33344	38801	32035	39553	37935	39021	42078	31370	34625	41019	43065	444443

There are two calls in January, one in April, one in July and one in September with type AA. The reason for these unknown (and unexisting) services is unclear to us, and we consider them as system bugs.

Note the increase in the fraction of calls of type "IN" (Internet technical support) over the year, and in particular the jump in July and August. This is seen more clearly in Figure 7. One explanation is the bank's expansion of Internet services – as the customer base widened, so did the need for technical support. Another possible reason could be changes into more complex hardware/software that affect Internet users. Figure 7 is similar to Figure 2, but for the different types of service. Type PS has not been included in the figure since it has a very similar pattern to the overall number of calls, and it dominates the other types (close to 70% of total).

Figure 7: Number of calls per month according to types



For staffing support, it is important to graph the number of calls according to types, since different agents are trained to provide different types of services. Along these lines, Figure 24 (p. 58) is similar to Figure 4 but is split according to types. One can see there the very clear peaks for type NE at times of opening and closing of the stock market, and then the very sharp decrease after the second peak. Type IN has peaks around 18:00h (possibly customers returning home after work, and then connecting) and around 22:00h (in Israel, this is the time at which reduced telephone rates go into effect).

Table 11: Call counts by priorities

Pr	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Total
0	18224	19602	15490	13880	17917	20076	17221	23112	21086	26103	21146	21311	235168
	(57.7)	(58.8)	(39.9)	(43.3)	(45.3)	(52.9)	(44.1)	(54.9)	(67.2)	(75.4)	(51.6)	(49.5)	(52.9)
1	4530	4623	7844	6274	7164	5666	7583	6573	3626	3203	6804	7937	71827
	(14.3)	(13.9)	(20.2)	(19.6)	(18.1)	(14.9)	(19.4)	(15.6)	(11.6)	(9.3)	(16.6)	(18.4)	(16.2)
2	8845	9119	15467	11882	14472	12193	14218	12393	6659	5319	13069	13817	137453
	(28)	(27.3)	(39.9)	(37.1)	(36.6)	(32.1)	(36.4)	(29.5)	(21.2)	(15.4)	(31.9)	(32.1)	(30.9)
Tot	31599	33344	38801	32036	39553	37935	39022	42078	31371	34625	41019	43065	444448

Table 12 divides the calls according to priority and type of service. Since customers who were transfered to an agent directly from the VRU, were recorded as having ID = 0 and priority 0, we split the priority 0 customers in services PS and NE to those with ID = 0 (hence their priority is actually unknown), and those with ID $\neq$ 0 for which it is known. We denote the customers with ID = 0 by "PS, -" and "NE, -". Hence "PS, 0" is used for customers with real 0 priority. We did not do this separation of priority 0 for types PE and TT, since the number of calls there is much smaller.

Table 12: Call counts by types and priorities

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Total
PS, -	11268	10229	6900	7526	10852	11929	8802	10781	12015	15889	10949	10533	127673
	(35.7)	(30)	(17.8)	(23.5)	(27.4)	(31.4)	(22.6)	(25.6)	(38.3)	(45.9)	(26.7)	(24.5)	(28.7)
PS, 0	32	30	24	31	65	20	30	70	13	15	28	97	455
	(0.1)	(0.1)	(0.1)	(0.1)	(0.2)	(0.1)	(0.1)	(0.2)	(-)	(-)	(0.1)	(0.2)	(0.1)
PS, 1	3956	3915	6591	5363	6172	4992	6678	5907	3181	2840	5876	6891	62362
	(12.5)	(11.7)	(17)	(16.7)	(15.6)	(13.2)	(17.1)	(14)	(10.1)	(8.2)	(14.3)	(16)	(14)
PS, 2	7742	7286	12395	9631	11393	9667	11759	10554	5601	4401	10303	11300	112032
	(24.5)	(21.9)	(31.9)	(30.1)	(28.8)	(25.5)	(30.1)	(25.1)	(17.9)	(12.7)	(25.1)	(26.2)	(25.2)
NW, 0	5562	7539	6701	4734	5098	5808	5699	6882	4323	5213	5090	4765	67414
	(17.6)	(22.6)	(17.3)	(14.8)	(12.9)	(15.3)	(14.6)	(16.4)	(13.8)	(15.1)	(12.4)	(11.1)	(15.2)
NE, -	740	1107	731	714	1039	1374	1045	1381	1389	1729	1369	1338	13956
	(2.3)	(3.3)	(1.9)	(2.2)	(2.6)	(3.6)	(2.7)	(3.3)	(4.4)	(5)	(3.3)	(3.1)	(3.1)
NE, 1	147	256	511	351	480	313	476	382	315	262	624	692	4809
	(0.5)	(0.8)	(1.3)	(1.1)	(1.2)	(0.8)	(1.2)	(0.9)	(1)	(0.8)	(1.5)	(1.6)	(1.1)
NE, 2	699	1387	2248	1711	2492	2089	2015	1618	931	824	2425	2138	20577
	(2.2)	(4.2)	(5.8)	(5.3)	(6.3)	(5.5)	(5.2)	(3.8)	(3)	(2.4)	(5.9)	(5)	(4.6)
IN, 0	295	293	502	451	492	483	1201	3562	3139	2987	3257	4067	20729
	(0.9)	(0.9)	(1.3)	(1.4)	(1.2)	(1.3)	(3.1)	(8.5)	(10)	(8.6)	(7.9)	(9.4)	(4.7)
PE, 0	79	65	48	36	59	74	48	72	47	68	61	55	712
	(0.3)	(0.2)	(0.1)	(0.1)	(0.1)	(0.2)	(0.1)	(0.2)	(0.1)	(0.2)	(0.1)	(0.1)	(0.2)
PE, 1	38	32	50	42	47	48	69	77	27	35	47	43	555
	(0.1)	(0.1)	(0.1)	(0.1)	(0.1)	(0.1)	(0.2)	(0.2)	(0.1)	(0.1)	(0.1)	(0.1)	(0.1)
PE, 2	47	52	65	39	40	34	43	32	39	38	62	66	557
	(0.1)	(0.2)	(0.2)	(0.1)	(0.1)	(0.1)	(0.1)	(0.1)	(0.1)	(0.1)	(0.2)	(0.2)	(0.1)
TT, 0	246	339	584	387	312	388	395	364	159	202	392	456	4224
	(0.8)	(1)	(1.5)	(1.2)	(0.8)	(1)	(1)	(0.9)	(0.5)	(0.6)	(1)	(1.1)	(0.9)
TT, 1	379	386	679	507	445	295	351	183	92	63	250	305	3935
	(1.2)	(1.2)	(1.7)	(1.6)	(1.1)	(0.8)	(0.9)	(0.4)	(0.3)	(0.2)	(0.6)	(0.7)	(0.9)
TT, 2	347	363	740	486	539	397	385	171	82	50	272	304	4136
	(1.1)	(1.1)	(1.9)	(1.5)	(1.4)	(1)	(1)	(0.4)	(0.3)	(0.1)	(0.7)	(0.7)	(0.9)
Tot	31577	33279	38769	32009	39525	37911	38996	42036	31353	34616	41005	43050	444126

There were 3 calls in July and one in September with "NE, 0", so we omitted this category from the table. It turns out that IN customers are not subject to identification requirements. In fact, they need not even have a bank account. Anyone can call and get instructions, via a dedicated number, on how to connect to the bank's site. Hence, beside 3 phone calls during the year with priority 1 or 2, all IN calls had priority 0.

There were 314 calls over the year with service type NW, ID $\neq$ 0 and priority different than 0. Since the type of service is identified by the phone number to which a customer calls, we do not understand how this is possible (there is no identification mechanism for new customers), and we consider such calls as system bugs, hence we did not include type NW with priorities 1 and 2 in the above table.

4.3 On the arrival process and Queueing Theory

Our empirical descriptions of the arrival process have been in terms of average arrival rates, as they vary over time. We refine the present descriptions in Section 9, where we focus on a single day (operational level). At the daily level we decompose the arrival process into its building blocks, that correspond to service types: PS, IN, NE and NW. All these descriptions are deterministic ("fluid-like") descriptions in terms of averages, which capture predictable variability. They are practically sufficient at the strategic and tactical level. However, at the operational level they aggregate too much stochastic variability for it to be ignored, and indeed, the operational level is where Queueing Theory is applied. (See the intoduction to Section 9 for some elaboration.) The main goal of queueing models is to predict performance measures, such as waiting times, abandonment and agents' utilization, as they vary with given variability (predictable and stochastic), mainly within arrivals, services and customers' patience.

Arrivals to call centers are typically random. For our purposes, *randomness* can be explained as follows: there are many potential, statistically identical callers to the call center; there is a very small yet non-negligible probability for each of them calling at any given minute, say, and they decide on whether to call independently of each other. Under such circumstances, theory dictates that the arrival process fits well, what is called, a *Poisson process*.

The Poisson process is completely characterized by its arrival-rate function, as in Figure 4. The variability in the figure is a consequence of changes in the elements of the story above, which occur predictably over time. For example, more customers are likely to call at 10:30am than at 1:00pm (or perhaps the same 1:00pm customers are simply more likely to call at 10:30am). To emphasize this predictable variability over time, Figure 4 characterizes a Poisson process that is *said* to be *inhomogeneous* in time. A time-homogeneous Poisson process, or simply a Poisson process, is one whose arrival rate is constant over time. Common call-center practice is to assume constant arrival rates for, say, individual hours or half-hours. Such an approximation, by a piecewise-constant arrival-rate function, is reasonable if predictable variability does not change abruptly.

There are various statistical procedures for testing whether an arrival process adheres to the Poisson model, and if so, infer its arrival-rate function. We plan it as part of our future research agenda, where we shall be analyzing the overall process, as well as the arrival processes for each service type separately. The outcome of these tests is not obvious, as reality often violates the Poisson assumptions. An example is frequent redialing soon after abandonment (relevant here), or arrivals that depend on the state of the call center (not relevant to our call center - this could happen within a network of call centers, which are interconnected by centralized load-balancing.)

5 VRU time: starting the service process

In this section (and in Sections 6–7), time is measured in seconds, and its descriptive statistics are rounded off to the nearest second. In tables with monthly data, we added a last column labeled Ann, which summarizes the data annually. Throughout the document, histograms and graphs are displayed so that rare large observations do not obscure the shape of the main bulk of the data. Graphs are drawn for all of 1999, unless specified otherwise.

All inbound calls must pass through the Voice Response Unit, or VRU. (Outbound TT calls do not.) In this section we analyze the VRU-time, for calls that dialed a number destined for service by an agent.

Table 13 presents summary statistics for the time spent in VRU. Some of the calls have very long VRU times (heavy tail). Table 14 repeats Table 13 when considering only calls with positive VRU shorter than 60 seconds (which captures over 96% of the calls in Table 13). (We are unclear why there are negative VRU times). Figure 8 shows the histogram of the VRU time.

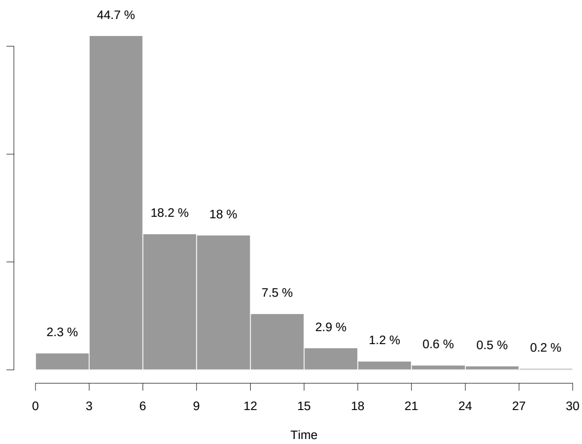
Table 13: Statistics for VRU time

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Ann
Mean	10	10	10	10	10	10	10	12	11	11	10	10	10
Med	9	9	6	6	6	8	6	9	9	9	6	6	8
SD	25	26	41	41	40	27	37	44	20	21	39	40	35
Min	-139	-192	-337	-210	-348	-210	-264	-362	-112	-71	-215	-341	-362
Max	1578	1860	2659	1922	3639	1925	1799	3625	1294	2168	4832	3257	4832
Tot	31599	33344	38801	32036	39553	37935	39022	42078	31371	34625	41019	43065	444448

Table 14: Statistics for VRU time, truncated at ± 60

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Ann
Mean	9	9	8	8	8	8	8	9	10	10	9	8	9
Med	9	9	6	6	6	8	6	8	9	9	7	6	8
SD	5	5	5	6	5	5	5	5	5	5	5	5	5
Tot	30665	32325	37349	30813	38336	36977	37817	40478	30815	34244	39964	41813	431596

Figure 8: Distribution of VRU time (1999)



6 Queueing time: waiting for service or abandoning

After the VRU experience, most customers move on to the queue: if there happens to be an idle server with the proper skills, service starts immediately and waiting time is 0; otherwise, the customer waits a positive amount of time until either service is granted or patience is lost (abandonment). In this section we analyze this queueing experience, starting with the aggregate, and then disecting it into its building blocks (separating the abandoning customers from those who got served.)

A central issue which, to the best of our knowledge, has never yet been addressed systematically, is the statistical nature of human patience on the phone. As will be explained in the sequel, the inference of patience must be based on censored data, which raises some statistical challenges. ([35] provides more details, accompanied by a short tutorial on censored sampling.) Our analysis also gives rise to three unexplored issues, the understanding of which requires inputs from psychology: first, the lack of a plausible definition of (im)patience (is a customer considered patient if willing to wait 5 minutes for service? the answer seems to be "yes" if the anticipated wait is, say, 1 minute, and "no" if it is 20 minutes); second, the need to distinguish between patience and "loyalty" or "persistency" (an impatient customer would still wait a long time for a badly-needed service); and third, the effect of information-while-waiting on patience (being reminded of one's wait allocates more attention to waiting which, in turn, increases impatience).

6.1 Time in queue: zero or positive

Table 4 (p. 12) reports how many of the calls had a positive waiting time, as opposed to those that did not wait. The latter occurs when customers are transfered directly to an AGENT from the VRU, or if they abandon from the VRU. Table 15 is similar to 4 but does not consider the customers abandoning from the VRU (i.e. a call is considered in the table if the time in queue is 0 and the outcome is AGENT, or if the time in queue is positive.)

Table 15: Call counts by waiting status (No VRU abandonment)

Q	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Total
$=0$	13475	13066	8669	8202	10892	14157	9714	13010	16113	21050	14787	14001	157136
	(44.4)	(41.6)	(23.5)	(27.8)	(29.1)	(39.2)	(26.4)	(33.4)	(54.1)	(63.6)	(37.8)	(34.2)	(37.4)
>0	16905	18306	28166	21345	26522	21993	27047	25962	13667	12047	24376	26889	263225
	(55.6)	(58.4)	(76.5)	(72.2)	(70.9)	(60.8)	(73.6)	(66.6)	(45.9)	(36.4)	(62.2)	(65.8)	(62.6)
Tot	30380	31372	36835	29547	37414	36150	36761	38972	29780	33097	39163	40890	420361

Tables 16 gives summary statistics for the calls reported in Table 15. The distribution of the time in queue is skewed to the right (mean significantly larger than median), with some extreme observations (over 5 hours), which we believe could have happened when customers do not disconnect the call properly. Table 17 presents the same summary statistics when calls with waiting time longer than 15 minutes were considered as outliers and not included in the calculations. (Table 17 captures over 99.9% of the calls in Table 16). Note that the median is 0 in September and October. This means that, among those seeking an agent and not abandoning from the VRU, more than half transfered directly to an agent. For some explanation, observe that at these two months, either the number of calls reaching the center was relatively low (September), or service times were short (October, to be displayed in Tables 38 and 39 later.) The result is service level high above average.

Table 16: Summary statistics for time in queue

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Ann
Mean	38	44	92	79	69	53	78	77	39	28	63	74	62
Med	11	13	46	38	29	16	38	29	0	0	20	28	20
SD	65	70	125	108	98	216	105	119	76	154	103	114	122
Tot	30380	31372	36835	29547	37414	36150	36761	38972	29780	33097	39163	40890	420361

Table 17: Summary statistics for time in queue, truncated at 15 minutes

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Ann
Mean	38	44	91	79	68	52	77	76	39	27	63	73	62
Med	11	13	46	38	29	16	38	28	0	0	20	28	20
SD	60	70	121	104	95	81	101	111	73	61	99	108	96
Tot	30369	31371	36797	29535	37401	36143	36742	38930	29776	33082	39138	40859	420143

6.2 Waiting time (wait > 0)

Zero wait consitutes the ultimate *operational* success of the call center service. Positive wait is a manifestation of a mismatch between demand (customers' service requests) and supply (agents' availability.) At its extreme, such a wait results in a system failure, namely an abandoning customer.

Of primary interest therefore is the *waiting time*, which here stands for a *positive* time in the queue. The overall summary statistics of waiting time are important, and no less so their stratification. For example, one is interested in the relationship between wait and prioirity, with the hope that high-priority customers wait less. Not that obvious is the relationship between wait/patience and service type: who enjoys the most/least patience? Answers for these and similar questions should support operational decisions, for example priority design, or balancing abandonment with the high staffing level that is required to prevent it.

Table 18 gives summary statistics for the waiting time of all the calls that waited, regardless of their outcome (served or abandoned). The distribution is skewed to the right. Table 19 presents the same statistics when waiting time is truncated at 15 minutes (capturing over 99.8% of the calls reported in Table 18). As expected, truncation reduces the mean (slightly) and the standard deviation (significantly). In fact, the two become rather close after truncation, annually as well as across months.

Table 18: Summary statistics for waiting time (wait > 0)

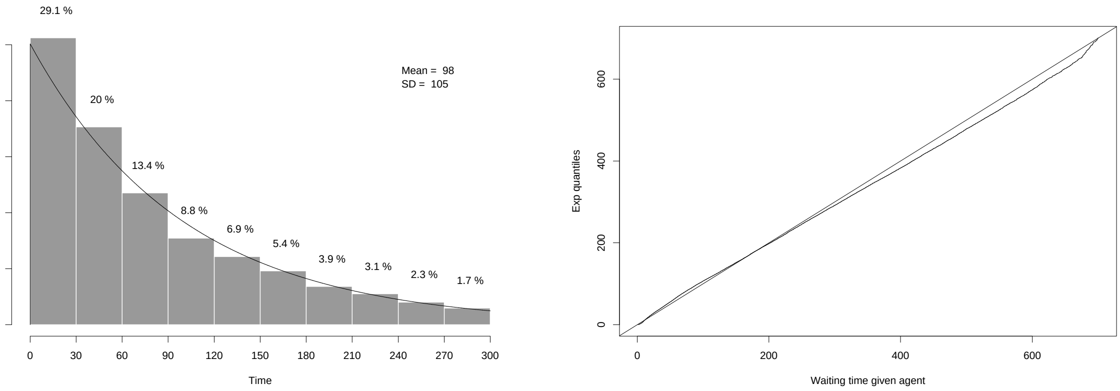
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Ann
Mean	68	76	120	110	97	88	106	116	85	76	102	113	100
Med	46	50	75	72	62	55	69	72	54	45	63	71	62
SD	75	78	130	113	104	271	110	130	93	247	114	125	142
Tot	16905	18306	28166	21345	26522	21993	27047	25962	13667	12047	24376	26889	263225

Table 19: Summary statistics for waiting time (wait > 0), truncated at 15 minutes

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Ann
Mean	68	76	119	109	96	85	105	114	84	73	101	111	98
Med	46	50	75	72	62	55	69	72	54	45	63	71	62
SD	67	78	126	108	101	89	105	119	89	83	109	116	105
Tot	16894	18305	28128	21333	26509	21986	27028	25920	13663	12032	24351	26858	263007

Figure 9 shows the waiting time histogram, which resembles that of the exponential distribution. This is consistent with the proximity of the mean and the standard deviation, that is a property of the exponential distribution. Another parametric verification is the ratio of mean to median which, for the exponential distribution, equals $\ln 2 = 0.69$, not far from the obeserved $62/98 = 0.63$. For a deeper distributional check, we superimposed the exponential density on top of the histogram, and created a Q-Q plot comparing the quantiles of waiting times to those of the exponential (the straight line at the right plot). The fit is reasonable up to about 700 seconds. (The p-value for the Kolmogorov-Smirnov test for Exponentiality is however 0 – not that surprising in view of the sample size of 263,007).

Figure 9: Distribution of waiting time (1999)



Remark on mixtures of independent exponentials: Interestingly, the means and standard deviations in Table 19 are rather close, both annually and across *all* months. This suggests also an exponential distribution for each month separately, as was indeed verified, and which is apparently inconsistent with the observerd annual exponentiality. The phenomenon recurs later as well, hence an explanation is in order. We shall be satisfied with demonstrating that a true mixture W of independent random varibles W_i, all of which have coefficients of variation $C(W_i) = 1$, can also have $C(W) \approx 1$. To this end, let W_i denote the waiting time in month i, and suppose it is exponentially distributed with mean m_i. Assume that the months are independent and let p_i be the fraction of calls performed in month i (out of the yearly total). If W denotes the mixture of these exponentials ($W = W_i$ with probability p_i, that is W has a hyper-exponential distribution), then

$$C^2(W) = 1 + 2C^2(M),$$

where M stands for a fictitious random variable, defined to be equal m_i with probability p_i. One concludes that if the m_i's do not vary much relative to their mean ($C(M) << 1$), which is the case here, then $C(W) \approx 1$, allowing for approximate exponentiality of both the mixture and its constituents.

6.2.1 The various waiting times, and their ramifications

We first distinguished between queueing time and waiting time. The latter does not account for zero-waits, and it is more relevant for managers, especially when considered jointly with the fraction of customers that did wait. A more fundamental distinction is between the waiting times of customer that got served and those that abandoned. Here is it important to recognize that the latter does *not* describe customers' patience, which we now explain.

A third distinction is between the time that a customer *needs* to wait before reaching an agent vs. the time that a customer is *willing* to wait before abandoning the system. The former is referred to as *virtual waiting time*, since it amounts to the time that a (virtual) customer, equipped with an infinite patience, would have waited till being served; the latter will serve as our operational measure of customers' patience. While both measures are obviously of great importance, note however that neither is directly observable, and hence must be estimated.

For customers who are patient enough to reach an agent, their waiting time is a sample of the time needed to wait, but it is only a lower bound for the time customers are willing to wait. In this case we say that the customer's patience (time willing to wait) is *censored* by the waiting time for service (time needs to wait), where only the latter is directly observable. And vice versa for customers who abandon, in which case the observable is the time that they are willing to wait, which censors the time needed to wait. The statistical branch that is devoted to censored data is *Survival Analysis*. Its basic mission is to develop tools for "uncensoring" data, for example to estimate customers' patience.

Naive censoring – the simplified MLE approach: We argue that waiting times of customers that were served or abandoned, both within types/priorities and across them, ought to play a central role in the analysis of a call center. As will be now demonstrated, their proper use and interpretations require both machinery and care. We start with summarizing some relevant findings from Tables 20 and 22 (these tables repeat Table 18 but with the additional segmentation according to call outcome – AGENT or HANG). There were 196,007 (75.5%) delayed customers who reached an agent with an average waiting time of 105 seconds, and 63,619 (24.5%) customers abandoning within 79 seconds, on average. What does this say about the time, R, that customers are willing to wait (patience) and the time, V, that they must wait (virtual wait)? About 75% lucky customers encountered $V < R$, while for the others $V > R$ and hence they abandoned. How does one infer the characteristics of V and R from the observed $W = min\{R, V\}$? For tutorial sake, assume that both V and R are exponentially distributed. Then under sufficient independence, the Maximum Likelihood Estimators (MLE) for $E(R)$ and $E(V)$ are given by [27]:

$$\begin{aligned} E(R) &= 79 + 105 \times \frac{75.5}{24.5} = 403 \quad \text{seconds,} \\ E(V) &= 105 + 79 \times \frac{24.5}{75.5} = 131 \quad \text{seconds.} \end{aligned}$$

While the 79 and 105 figures were observed (mean waiting time among customers who waited and abandon and among customers who waited and reached an agent, respectively), it is the inferred 403 and 131 seconds that are of managerial significance.

One can give the MLE an intuitive justification as follows. Consider R for example. Then think of waiting customers as tossing a coin *every second*, in order to decide on whether to abandon or remain in queue for the next second. Our data indicates that overall there were

$$63,619 \times 79 + 196,007 \times 105 = 25,606,636$$

such tosses, out of which $63,619$ ended up with an abandonment. The probability that a single toss results in abandonment is, therefore, estimated to be the observed fraction of abandonment $63,619/25,606,636$. Its reciprocal is the average number of tosses (seconds) till abandoning, namely $25,606,636/63,619 \approx 403$, as derived above. (In other words, since one out of 403 tosses results in abandonment, on average, and the coin is tossed every second, then customers' average patience is estimated to be 403 seconds.) In this demonstration, we assumed the Exponential distribution. Using survival analysis tools allow us to obtain estimates for $E(R)$ and $E(V)$ without making such parametric assumptions.

The conclusion is that customers who are familiar with the system seem to be patient: in order to get served, they are willing to wait more than 3 times of what they expect to wait. (The estimated ratio between $E(R)$ and $E(V)$ is actually higher – about 5 to 1, as follows from a non-parametric analysis that is summarized in Tables 36 and 37.) One could similarly analyze waits for customer types and priorities and then compare their patience. This will be carried out in Section 6.3. For now, and throughout the rest of the subsection, we segregate our waiting time data into those that got served and the others who abandoned.

Tables 20 and 21 repeat Tables 18 and 19 for waiting times of abandoning customers. (99.7% of the calls counted in Table 20 were captured in Table 21). Tables 22 and 23 repeat Tables 18 and 19 for customers that reached an agent (99.8% of the calls counted in Table 22 were captured in Table 23). Figures 10 shows the histogram of waiting time given abandonment.

Table 20: Waiting time when abandon

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Ann
Mean	61	66	85	85	72	71	82	90	76	79	80	86	79
Med	39	43	56	56	43	47	53	54	51	49	52	56	51
SD	95	73	101	104	91	80	103	129	101	112	100	120	104
Tot	3080	3932	8581	5659	7073	4635	6787	7260	2555	2089	5495	6473	63619

Table 21: Waiting time when abandon, truncated at 15 min

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Ann
Mean	57	66	84	83	70	71	80	85	74	77	79	84	78
Med	39	43	55	56	43	47	53	54	51	49	52	55	51
SD	66	71	96	90	84	79	91	101	81	95	94	98	90
Tot	3071	3931	8573	5651	7066	4634	6780	7237	2553	2084	5491	6460	63531

Table 22: Waiting time when reaching an agent

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Ann
Mean	70	79	135	118	105	89	113	126	86	73	107	120	105
Med	48	52	91	83	71	58	78	83	55	44	68	80	67
SD	67	79	138	114	105	91	110	128	91	87	116	125	111
Tot	13585	14096	19178	15381	19031	17036	19882	18392	10949	9811	18584	20082	196007

Table 23: Waiting time when reaching an agent, truncated at 15 minutes

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Ann
Mean	69	79	134	118	105	89	112	125	86	72	106	119	105
Med	48	52	91	83	71	58	77	83	55	44	68	80	67
SD	66	79	133	112	104	91	108	124	90	80	112	119	108
Tot	13584	14090	19150	15377	19027	17033	19871	18374	10947	9802	18565	20064	195890

Figure 11 shows the histogram of waiting time given served (reaching an agent). Note that the means and SD are close to each other, which again suggests exponentiality. Comparing the empirical distribution to the Exponential density, and to the Exponential quantiles using Q-Q plot (right plot), yields a fit that is even better than before. (The Kolmogorov-Smirnov test still has a p-value of 0.) The phenomenon of exponential mixtures from Table 19 is even sharper here. It arises across months, as before, but also in getting Table 19 as a mixture of Table 21 (25% of the observations) and Table 23 (75%).

Figure 10: Distribution of waiting time given abandonment

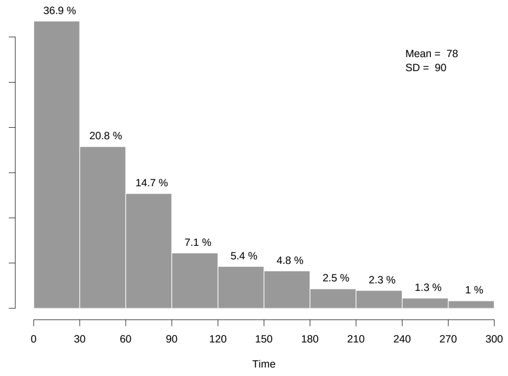
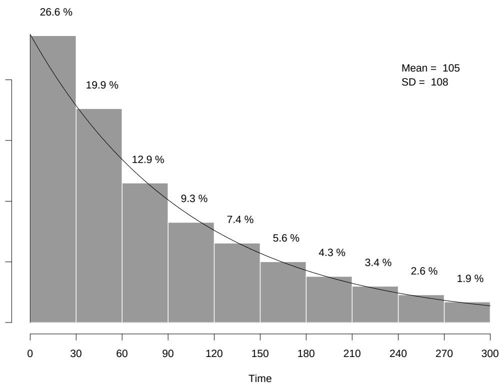
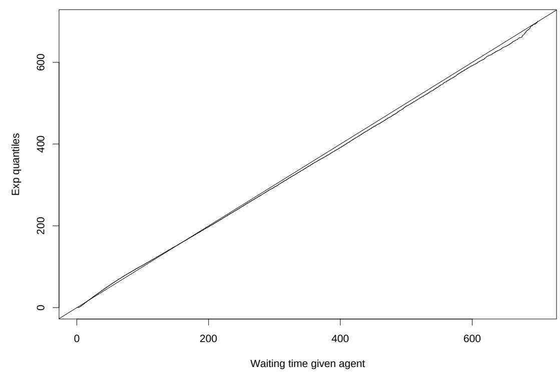


Figure 11: Distribution of waiting time given reaching an agent





We conclude the Subsection with refinements of Tables 20–23, which account for the four main types of service (PS, NE, IN and NW) and for priorities 1 and 2. This is summarized in Tables 24–35, in which the waiting times were all truncated at 15 minutes.

Table 24: Waiting time when abandoning for type PS

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Ann
Mean	51	53	69	69	54	55	62	67	57	54	61	64	62
Med	37	37	50	50	33	37	43	48	41	39	43	47	43
SD	61	57	76	72	68	61	67	74	57	60	67	69	69
Tot	1643	1666	4880	3340	4600	2439	3894	3585	1192	735	2935	3657	34566

Table 25: Waiting time when abandoning for type NW

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Ann
Mean	58	71	95	97	96	86	101	90	87	70	87	97	88
Med	40	47	61	64	62	60	63	56	58	46	56	63	58
SD	61	74	100	93	96	85	108	102	94	79	95	101	94
Tot	1203	1950	2810	1778	1801	1713	1979	2363	946	906	1505	1620	20574

Table 26: Waiting time when abandoning for type IN

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Ann
Mean	135	143	172	178	132	151	133	142	101	138	143	140	140
Med	101	102	92	119	76	86	93	84	71	85	82	86	86
SD	112	123	191	170	124	160	125	153	99	147	153	152	148
Tot	58	62	208	159	170	124	411	942	286	345	531	713	4009

Table 27: Waiting time when abandoning for type NE

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Ann
Mean	108	91	121	116	100	73	87	87	79	70	98	107	99
Med	58	53	75	69	53	40	50	50	48	44	56	72	55
SD	113	94	133	136	116	84	99	101	71	76	111	117	113
Tot	83	181	496	283	398	285	382	243	80	57	386	346	3220

Table 28: Waiting time when abandoning for priority 1 customers

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Ann
Mean	57	62	74	72	65	59	70	71	63	64	67	71	68
Med	43	44	54	52	43	40	50	53	46	48	49	52	49
SD	63	66	77	80	75	66	73	76	62	71	73	76	74
Tot	798	911	2451	1752	1971	1197	1902	1789	657	440	1579	1990	17437

Table 29: Waiting time when abandoning for priority 2 customers

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Ann
Mean	54	56	74	72	67	60	66	67	55	49	65	67	66
Med	35	38	49	48	43	40	44	46	38	33	41	45	44
SD	71	62	89	79	80	65	76	78	60	58	77	80	77
Tot	920	902	3034	1911	2160	1401	2175	2022	643	365	1822	2066	19421

Table 30: Waiting time when reaching and agent for type PS

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Ann
Mean	65	69	126	110	98	80	104	112	78	60	98	107	96
Med	45	47	84	77	67	53	72	76	50	39	63	72	62
SD	61	67	128	104	97	80	99	109	80	60	101	105	98
Tot	10038	9594	13887	11486	14219	12305	14787	12885	7494	6426	13062	14383	140566

Table 31: Waiting time when reaching and agent for type NW

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Ann
Mean	81	107	169	152	147	122	165	168	119	97	132	171	136
Med	55	73	125	117	109	79	125	123	74	60	84	127	92
SD	77	101	152	133	131	119	143	147	120	101	127	151	131
Tot	2337	2577	2402	1679	1770	2115	2231	2254	1386	1477	1688	1671	23587

Table 32: Waiting time when reaching an agent for type IN

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Ann
Mean	124	145	256	219	189	176	159	193	115	135	147	156	159
Med	89	112	225	181	147	147	113	140	78	87	85	92	103
SD	109	134	196	166	146	151	142	173	117	138	165	167	159
Tot	128	139	150	139	167	201	423	1292	763	753	898	1265	6318

Table 33: Waiting time when reaching and agent for type NE

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Ann
Mean	82	88	147	138	115	97	110	117	83	66	122	137	114
Med	56	61	109	100	82	69	80	76	57	44	84	100	78
SD	78	85	132	128	108	95	101	115	79	66	123	128	112
Tot	762	1456	2251	1781	2557	2107	2118	1752	1165	1026	2645	2471	22091

Table 34: Waiting time when reaching and agent for priority 1 customers

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Ann
Mean	69	81	145	129	117	97	117	130	94	74	118	131	113
Med	48	58	108	100	86	66	90	98	65	49	85	96	79
SD	63	75	136	115	106	91	104	118	90	70	113	119	108
Tot	3395	3327	4765	3997	4709	4143	5292	4527	2842	2655	4938	5583	50173

Table 35: Waiting time when reaching and agent for priority 2 customers

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Ann
Mean	65	67	121	106	95	78	99	105	71	54	94	102	92
Med	45	45	80	72	65	51	68	69	47	37	60	69	60
SD	62	67	124	103	96	79	96	106	73	54	101	102	96
Tot	7586	7864	11682	9455	11771	10383	11631	10100	5905	4867	10944	11376	113564

6.3 Hazard rates and survival functions for patience and virtual wait

In this subsection, we apply survival analysis [27, 22] to help us understand the various times in queue. Two outcomes are possible for a customer that joins the queue: AGENT or HANG. (We ignore the rare PHANTOM outcomes). If one is interested in "how much time a customer is *willing* to wait before abandoning" then a customer who reached an agent is a censored observation. If the interest is in "how much time a customer *needs* to wait before reaching an agent" (virtual waiting time), then a customer who abandons is a censored observation. As before, denote by R the "time willing to wait", by V the "virtual waiting time", and equip both with their steady-state distributions. One actually samples $W = \min\{R, V\}$, as well as the indicator $1_{\{R<V\}}$ for observing R or V. To estimate the distribution of R, one considers all calls that reached an agent as censored observations, and vice versa for estimating the distribution of V.

The hazard rate function of R provides a natural dynamic depiction of patience, as it evolves while waiting. This was first recognized by Palm [28], who used it to define *irritation* as a function of the waiting time. The *empirical hazard* over a time interval is the number of failures during that interval (i.e. the number who abandon when R is of interest, and the number of served calls when V is), divided by the number at risk at the beginning of the interval (i.e. the number of calls who were still waiting at the beginning of the interval). These raw hazard samples are noisy, and they become unstable at the right tail, since the remaining population diminishes there. The raw hazard rates are the building blocks for the classical Kaplan-Meier estimator of the survival function. (See [27], or the Appendix in [35] for a tutorial.) In order to estimate a *smooth* hazard rate function, the following options are available:

1. Using nonparametric regression methods: smooth the raw hazard rates as a function of time. This includes methods such as LOWESS, SUPER-SMOOTHER, Kernel-based smoothing or splines. These methods are explained in Hastie and Tibshirani [18] and in Härdle [16]. Venables and Ripley [33] describe the implementation in S-Plus. The input to all the procedures are pairs (t_i, h_i), where h_i is the estimated hazard rate at time t_i, or more precisely over the time interval $(t_{i-1}, t_i]$.

2. HEFT (Hazard Estimation with Flexible Tails): Kooperberg, Stone and Truong [24] use cubic splines for estimating the log-hazard. The method also incorporates two additional log terms to fit the extreme tails in a reasonable way. The algorithm picks, adaptively, a model among a set of competing nested models, using the BIC criteria (while maintaining hierarchy of the terms in the model). The advantage of using HEFT over general nonparametric regression methods is that it was customized to deal with hazard rates, hence it gives a more needed attention to the tails. From our experience, tails of hazard rates that are estimated by HEFT tend to be well-behaved, in fact sometimes too much so as to obscure legitimate variability.

3. The standard way to estimate hazard in a smooth way is via regression. Among regression models, the Cox proportional hazard model [8] is probably the most popular. HARE (HAzard REgression) is an extension of HEFT, proposed by Kooperberg et al. [24], which estimates hazard rates smoothly, in the presence of covariates. The method is non-parametric, and is actually an extension of the Cox model. As a plan for future research, we intend to apply regression-based inference, for example to analyze covariates that affect patience.

We experimented with all the methods mentioned above. Visually, graphs by HEFT and by nonparametric regression methods are similar for the main body of the data, but they differ at the tails. Indeed, we chose to present plots obtained from HEFT since they seem to fit the tails better. All plots is this section are based on November and December data.

Figure 12 (p. 33) shows the smoothed hazard rates for R – the time a customer is willing to wait before abandoning. The bottom left plot (vertically) exhibits a smooth HEFT estimate of the hazard, superimposed on the raw hazard rates. We truncate the picture at 400 seconds because otherwise the interesting pattern at small times is being obscured. Note the arched patterns occurring for larger times. The reason for the arches is merely visual (and is not rooted in any biased data). To see that, note that close to the tail, the number of failures in a given interval is small (in our picture 0,1,2,3 or 4 abandonment). This number is the numerator of the raw hazard rate, which is divided by the number at risk to produce the displayed points. As time increases, the denominator (number at risk) decreases, hence the arching upward (except when 0 is the numerator). Thus, each separate arch is associated with one of the 5 values in the numerator. The arches become more distinguished as time increases, and the remaining population becomes small (however, for different types this happens at different times.) As a rule of thumb, we do not trust the hazard estimate when the arches pattern is clear. The other plots show the smoothed hazard rates, obtained by HEFT, for the types of service PS, NE and NW. For type PS we include the hazard estimate for priorities 1 and 2, as well as for all calls. For type NE, we superimpose the hazard estimate on the raw hazard, as an explanation for the narrow peaks. We do not present the results for NE segmented by priorities, since there were too few calls of priority 1 for the estimates to be trustable.

An important comment regarding the hazard estimate of R is that, unlike "traditional" applications, of survival analysis, here the fraction of censoring is very large (about 75% – see Tables 20 and 22), which means that *any* method one would use to estimate the hazard cannot be trusted towards the tails. The quantification of such mistrust ought to be a subject for future research.

6.3.1 On information while waiting, and the perception of time

Local peaks in the hazard rates manifest systematic tendency to abandon. The plots in Figures 12 (as well as for other months) clearly depict such peaks around 15 and 60 seconds. There are also occasional peaks at other multiples of 60. This suggested that some systematic phenomenon is lurking in the background during waiting. And indeed, upon joining the queue, and about every minute or so thereafter, customers are exposed to an automatic message. (In the pictures we presented, sometimes there are additional peaks, which appear when there is a peak in the 'cloud' of raw hazard. However, we have not seen systematic times in which the peaks appear, and they might be due to noise or the smoothing algorithm).

The message informs customers about their relative position in the queue, accompanied by the waiting time of the longest-waiting customer. (The system takes into account the fact that if there are N agents at work, then the larger the N the faster is the global service rate: to this end, customers in positions 1 to N are considered "first" in queue, positions $N+1$ to $2N$ are "second", etc.) The hazard rate function clearly demonstrates that the message causes customers to abandon, which seems to contrast the reason for having such a message at the first place. (Messages are usually conceived to reduce anxiety that is rooted in the uncertainty behind waiting at tele-queues. The hope typically is to reduce abandonment, but occasionally also to encourage it during periods of excessive congestion.) Peaked abandonment as a response to a message is consistent with prevalent psychological theories of time perception: the message "reminds" delayed customers of their waiting, thus increasing the attention resources that they allocate to their delay. The result is an acceleration of the subjective time flow (a second "becomes" a minute), thus resulting in an increased likelihood for abandonment. One could also associate with the peaks a rational behavior, but we do not dwell on that here. To test the effects of messages, if any, one could start with comparing hazard rates for abandonment under differing congestion conditions.

6.3.2 On dependence, or the violation of the classical assumptions in survival analysis

A basic assumption in survival analysis is that failure times and censoring times are independent (or at least non-informative of each other). Otherwise, estimation of the failure time distribution could become non-identifiable. In our data, such independence assumptions are violated in many ways. As a start, inherent dependence exists between the virtual waiting time encountered by, say, customer n, and the patience of

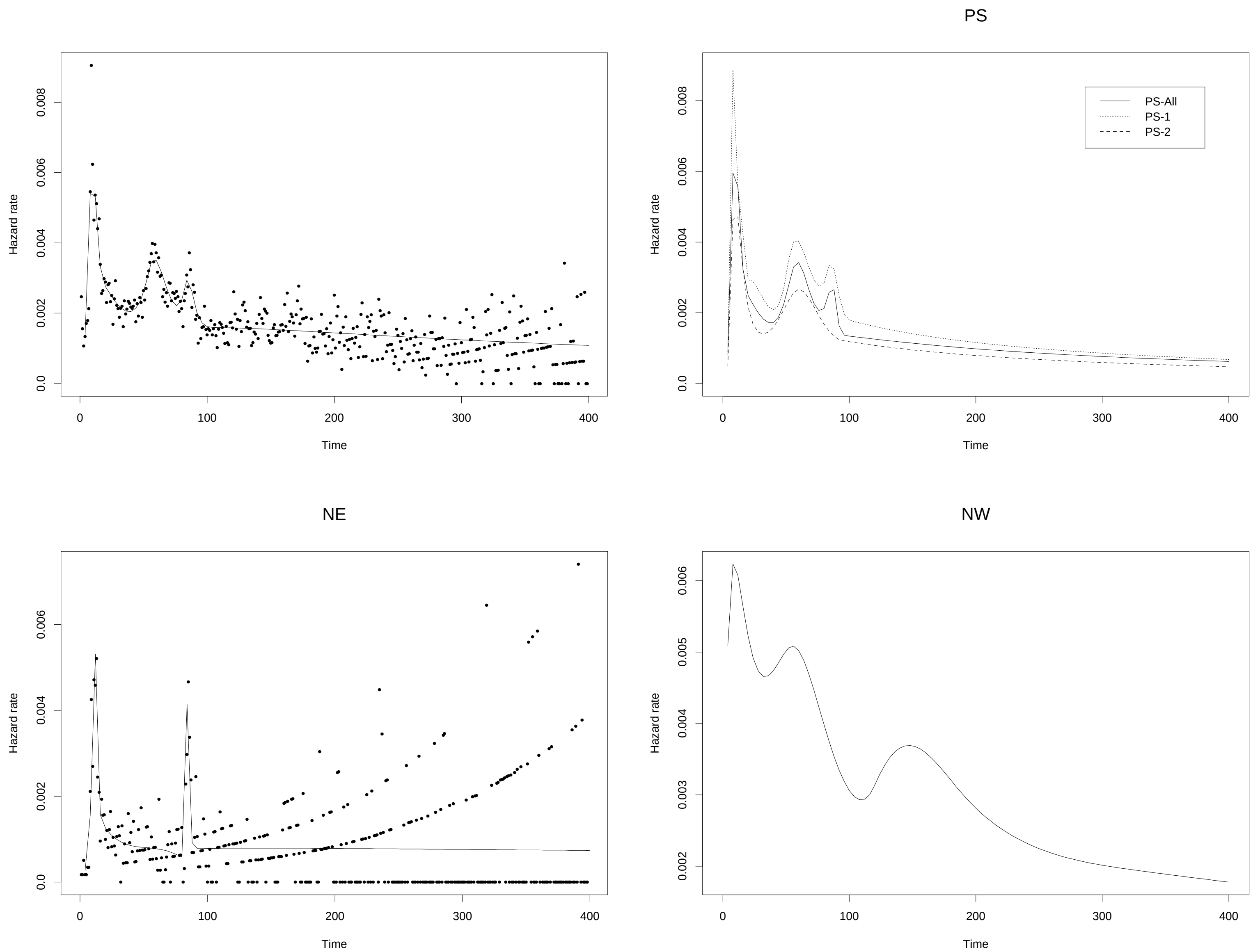


Figure 12: Hazard rate for time willing to wait (Nov + Dec)

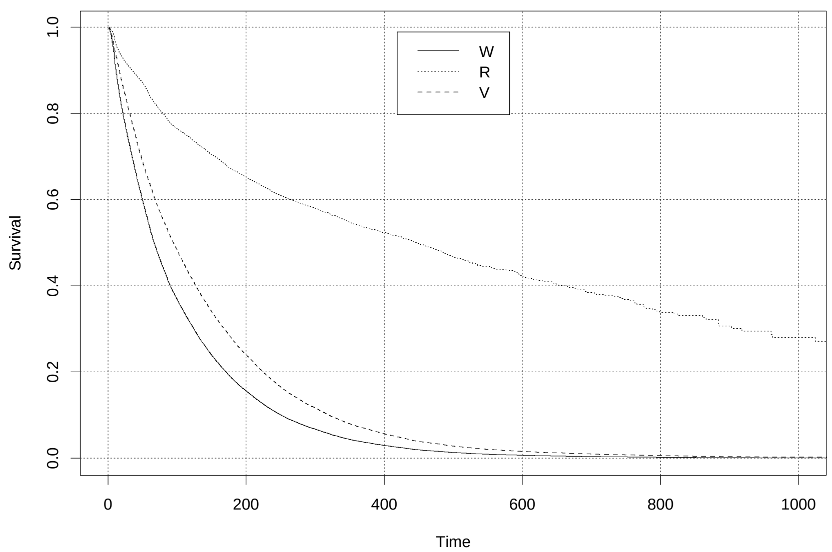
customer $n-1$ for example. Furthermore, a message as described above, informs customers about their queue, which possibly affects their patience. (Listening to the message ourselves, however, we found it to be rather confusing, hence its effects on customers, and the level of dependence that it introduces, remain unclear.) Another source for dependence is repeat calls by the same customers (sometimes many times during short periods). Finally, if the queue is long at some given time then the virtual waiting time is likely to be long as well; consequently both are likely to be long also soon thereafter. If follows that virtual waiting times are dependent across customers.

6.3.3 Patience and virtual wait across types and priorities

Having said the above, about the violations of the standard assumptions of survival analysis, it is important and useful to emphasize that our fitting of HEFT (or any other smoothing method for that matter), while ignoring all forms of dependence, has lead nevertheless to important and trustable insight. Consider, for example, Figure 13, where we plot the survival function of R (time willing to wait before abandoning), V (virtual waiting time) and $W = \min\{V, R\}$. A clear stochastic ordering emerges among the three distributions. Moreover, the same ordering arises at all months, and across different types of service. (Sometimes, however, the survival curve for R gets closer to that of V.) The reason for the survival function of W being the lowest is obvious: as their minimum, W is indeed smaller than both R and V. In contrast, the stochastic ordering between V and R is interesting and informative. It indicates that customers are willing to wait (R) more than they need to wait (V), which suggests that our customer population consists of patient customers. (Here we have implicitly, and only intuitively, defined a *patient customer*; systematic research on this subject is unfortunately lacking.)

The survival functions were estimated in two methods. The first via the classical Kaplan-Meier estimator, which is the empirical distribution function in the face of censoring. The second is via HEFT which was described above. With HEFT, we first estimated the hazard rate, then integrated it up to time t to get (after exponentiation) the survival function at time t. Both Kaplan-Meier and HEFT yield curves that are visually very close, at least for t over the ranges of interest to us (t not too large). We present the results

Figure 13: Survival curves (Nov + Dec)



obtained using Kaplan-Meier. Note that the survival curve reaches zero only if the largest observation is not censored.

In Figure 14 (p. 36) the hazard rates for the virtual waiting times are estimated for all calls, and according to types. The overall plot reveals rather remarkably constant behavior beyond 50 seconds, which suggests a memoryless (exponential) behavior of the tail. (This empirical exponentiality of the virtual wait should be further analyzed, especially in view of classical queueing models where it arises in great generality.)

In Figure 15 (p. 37), the survival functions of R are compared for different types, and then for types 'PS' and 'NE' according to priorities. Again, clear stochastic orderings emerge. First, we learn that 'NE' customers are willing to wait more then 'PS', for example. A possible explanation is that NE customers need the service more urgently. (One can not trade stocks tomorrow at today's prices.) Note that the plots are on different scales (for the lower two plots the y-scale is between 0.5 and 1, and not between 0 and 1). On the upper plot, all the survival curves reach 0 (at times larger than 800). However, the survival curves for type NE with priority 1, and for type PS with priority 1, do not reach 0. Actually, for NE with priority 1, the minimum value of the survival curve is 0.503, which will be elaborated on below (in our comment on estimating censored statistics).

Interestingly, the survival curve of PS with priority 1 is below that for PS with priority 2, but the picture is reversed for NE customers. As for PS customers, we are learning that higher priority customers are less likely to abandon at any given time during their wait. The behavior of NE customers is reversed. One would intuitively expect high priority customers to be less tolerant of delay. On the other hand, their need for service, and their trust of the system to provide it, might be higher. We are lead to distinguish, and trade off, between patience (or tolerance for waiting) and loyalty/persistency, which suggests yet an additional avenue for future research.

The plots in Figure 16 (p. 38) display the estimated survival function of V (virtual waiting time) for the different types, and for the two types 'PS' and 'NE' according to their priorities. Here the ordering among types is not that clean as with patience. With respect to priorities, however, the call center's manager would be relieved to learn that low-priority customers are in fact asked to wait more, but not much more. (This important information, regarding the amount of time that customers are required to wait, can not be deduced from direct observations – "uncensoring" must be carried out first.)

On estimating censored means, variances and medians: Means and variances of censored distributions have been estimated via the tail-formula, applied to the survival function. The mean of R is simply the integral of its survival function, say $\bar{F}(t)$, over $0 \le t < \infty$. For variance calculations, one uses the fact that the mean of R^2 is the integral of $2t \times \bar{F}(t)$.

The statistical problem of producing confidence intervals from censored data is generally difficult. In fact, the present study has stimulated research in this direction. For now, we content ourselves with some remarks on the accuracy of our procedures.

If the last observation in a censored data-set is in fact censored, then the estimated distribution is defective, with a positive mass at infinity. Consequently, the application of the tail formula, while omitting this last observation, results in a downward biased estimator. In our set-up, the last observation for either V or for R *must* be censored. Hence, depending on whether the outcome of the largest observation is AGENT or HANG, one of the two means is underestimated. If several observations are censored at the high tail, it may turn out impossible to estimate reliably even some of the quantiles. In particular, the median cannot be estimated reliably when the estimated survival curve does not reach 0.5. (This occurs for some of our data – see Figure 15.) Another scenario is when the curve does reach 0.5, but there are large gaps between observations (i.e. the width of the steps around 0.5 is large). We applied the tail formula to estimate the mean and SD through numerical integration. Testing our procedure, we calculated from survival functions the mean and SD of service times (where *no* censoring is called for). The resulting means were found very close to the sample means, and the SD's were slightly less accurate (and typically smaller than the sample SD).

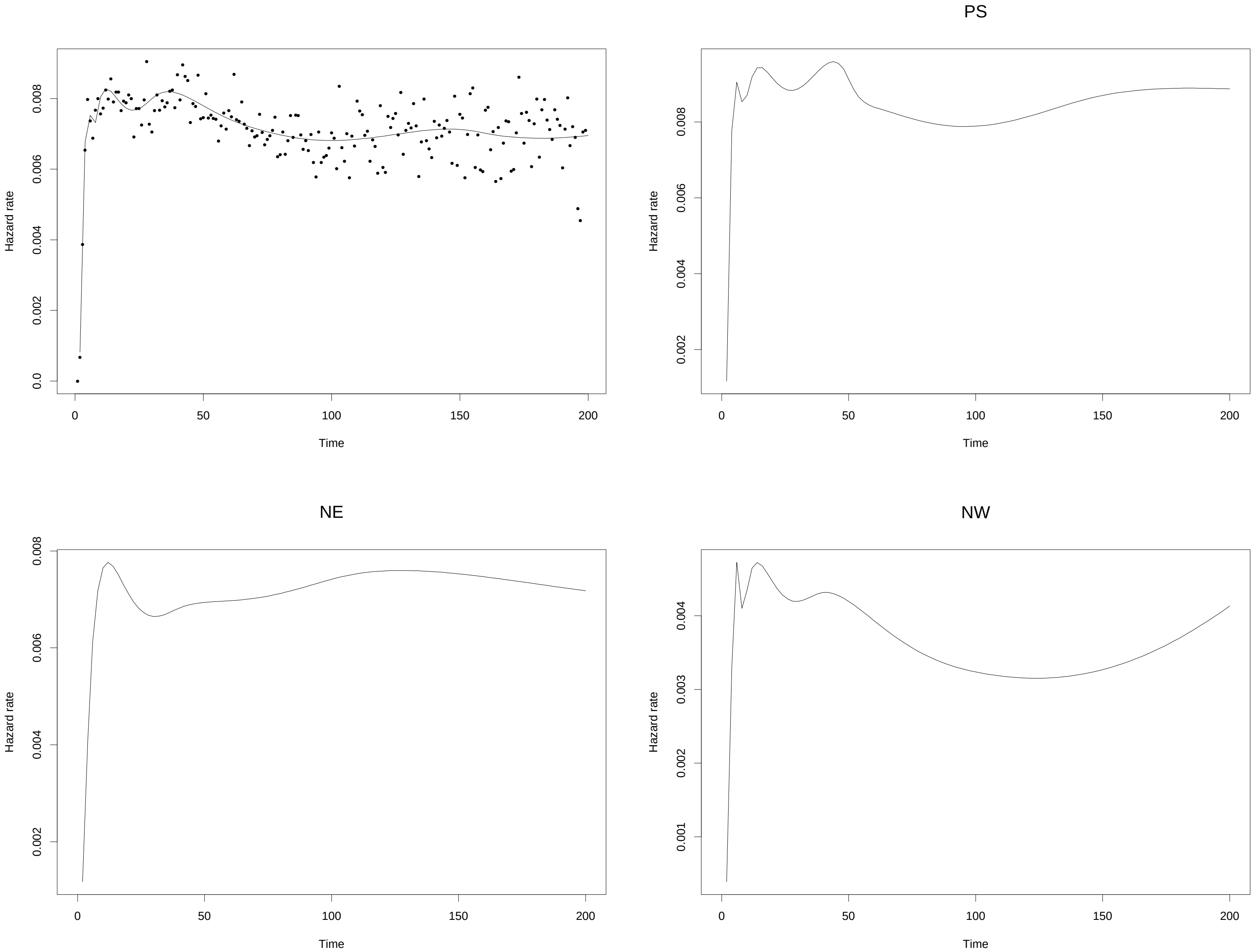


Figure 14: Hazard rates for virtual waiting time (Nov + Dec)

Figure 15: Survival curves for time willing to wait (Nov + Dec)

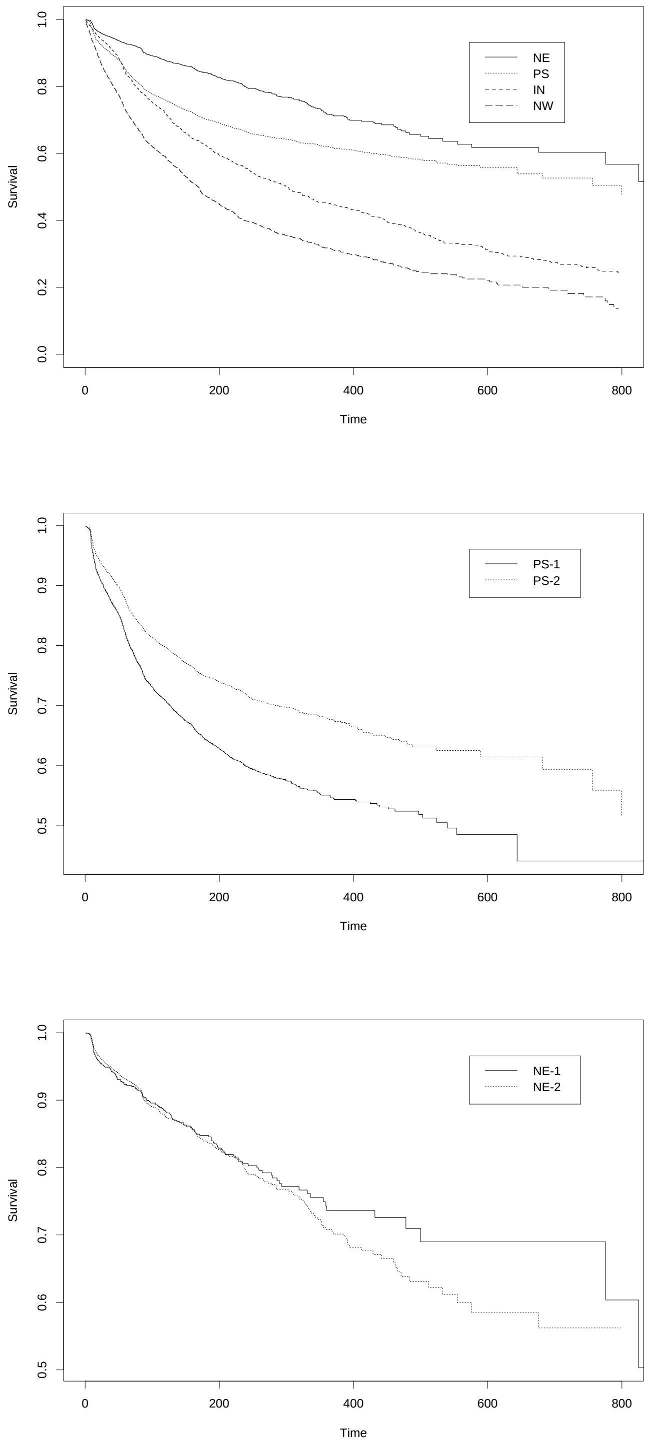
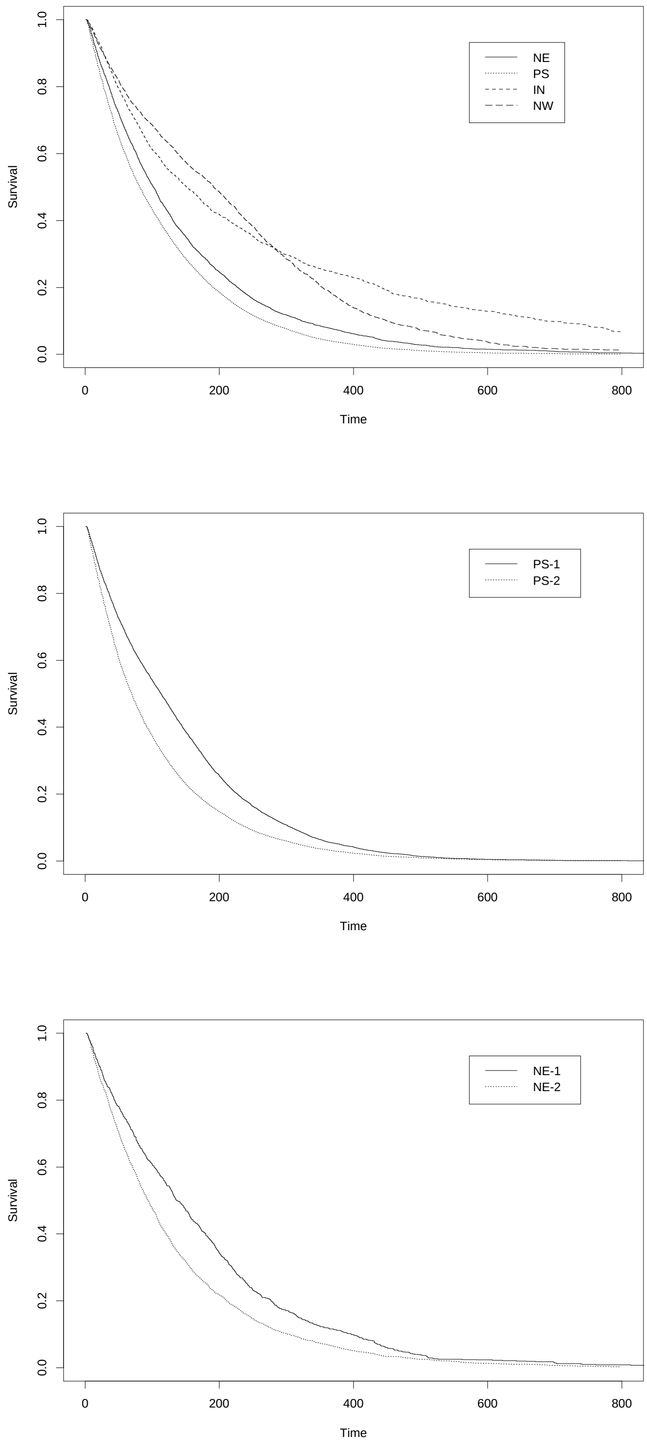


Figure 16: Survival curves for virtual waiting time (Nov + Dec)



Tables 36 and 37 (p. 39) summarizes the estimated mean and variance of V and R, for all calls as well as by stratification according to types and priorities for NE and PS. The findings confirm previous assessments on the overall patience of our customers, with refinements for the various types. Specifically, define the *patience index* to be the ratio between the time a customer is willing to wait and the time needed to wait. The overall patience index is 5.25, and for the various types it is as follows: PS = 5.10, NE = 4.74, NW = 2.14, IN = 1.98. (PS are found most patient and IN the least.)

As a final remark, recall that for either the mean of R or V, one of their estimators must be biased downwards, having a largest observation that is censored. Perhaps a more reliable estimate of location would be the median (taken to be the $50th$ percentile of the estimated survival curve). However, as already mentioned, problems arise here as well (for example see the median of R for types PS and NE – Table 36).

Table 36: Means and medians for V and R from Kaplan-Meier
(Nov + Dec)

		Mean	SD	Median
Time willing to wait (R):	General	741	805	446
	PS	597	409	782
	NE	678	362	979
	NW	491	777	168
	IN	528	560	301
Time needs to wait (V):	General	141	161	94
	PS	117	113	81
	NE	143	140	101
	NW	229	251	191
	IN	268	313	150

Table 37: Means and medians for V and R according to priorities
(Nov + Dec)

		Mean	SD	Median
Time willing to wait (R):	PS – General	597	409	782
	PS – priority 1	521	399	536
	PS – priority 2	644	396	981
	NE – General	678	362	979
	NE – priority 1	703	354	979
	NE – priority 2	647	349	934
Time needs to wait (V):	PS – General	117	113	81
	PS – priority 1	140	121	112
	PS – priority 2	103	106	68
	NE – General	143	140	101
	NE – priority 1	175	154	136
	NE – priority 2	133	133	93

7 Service time: ending the service process

The last phase in a successful visit to the call center is typically the service itself. Table 38 provides summary statistics for service times (ignoring records with zero service time). The distribution is skewed to the right. Table 39 presents the same statistics when considering only calls with service time shorter than 1 hour = 3600 seconds. (Table 39 captures over 99.9% of the calls from Table 38). Figure 17 displays the service time histograms: one histogram is for the first 10 months of 1999 (98% of the service times are within the range of the histogram), and the other is for November-December (here 97.6% are captured).

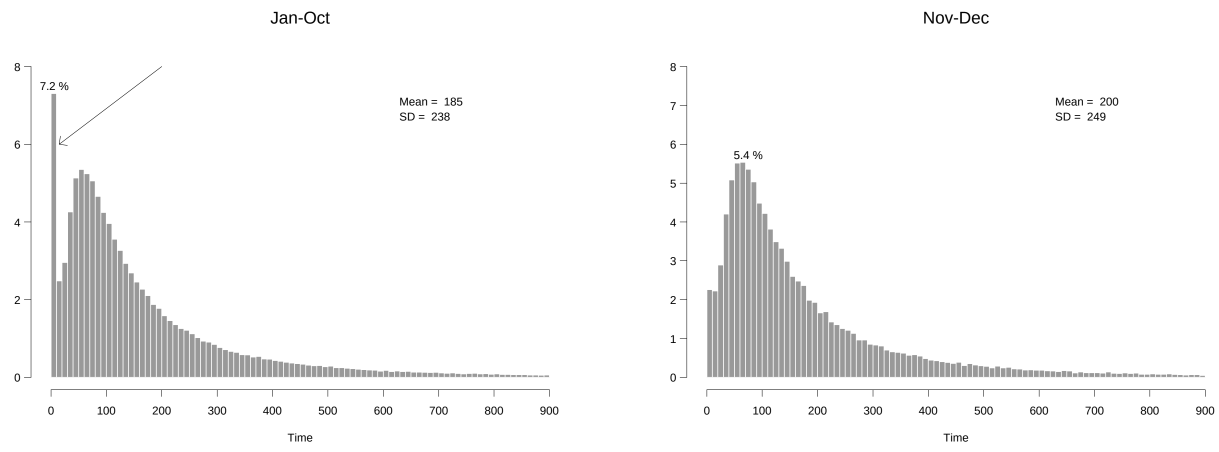
Table 38: Service time

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Ann
Mean	184	175	184	211	197	189	182	183	191	181	196	207	190
Med	113	106	110	123	117	115	112	110	111	105	117	128	114
SD	230	217	239	365	259	313	301	362	296	437	272	273	304
Tot	27091	27451	28278	23923	30403	31492	29927	31713	27160	31009	33708	34433	356588

Table 39: Service time, truncated at 1 hour of service (3600 seconds)

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Ann
Mean	183	175	183	207	196	185	179	177	189	177	194	205	188
Med	113	106	110	123	117	114	112	109	111	105	117	128	114
SD	224	213	232	260	250	232	226	229	264	245	249	249	240
Tot	27087	27448	28272	23916	30397	31479	29917	31680	27145	30994	33693	34419	356447

Figure 17: Distribution of service time



Note the high percentage of calls with service time shorter than 10 seconds during January–October. To see this more clearly, Table 40 summarizes the cumulative percentages of customers receiving service time shorter than the time at the column labeled “Time”. For example, 10% of the March-services lasted less than 15 seconds.

Table 40: Cumulative percentages for short Service time

Time	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1	0.4	0.7	0.7	0.7	0.6	1	1	0.7	0.9	0.6	0.003	0
2	0.8	1.5	1.8	1.5	1.8	2.3	2.2	1.6	2.3	1.5	0.04	0.01
3	1.7	2.8	3.2	2.7	3.4	3.4	3.6	2.7	3.5	2.2	0.2	0.3
4	2.6	4	4.4	3.9	4.8	4.4	4.7	3.6	4.2	3	0.7	0.7
5	3.3	4.8	5.4	4.8	5.8	5.3	5.6	4.2	4.8	3.5	1	1.1
10	5.7	7.5	8.4	7.6	8.8	7.6	8.1	6.3	6.6	5.1	2.2	2.2
15	6.8	9	10	8.9	10.5	8.9	9.2	7.4	7.6	6.2	3.3	3.2
20	7.8	10.4	11.2	9.9	12	9.9	10.3	8.5	8.6	7.3	4.5	4.3
30	10.7	13.8	14.3	12.3	15.2	12.8	12.8	11.2	11.3	10.4	7.5	6.9

While service times of 30 seconds are perhaps conceivable, service times of 5 or even 10 seconds, at the frequency encountered, are questionable. And indeed, questioning the manager of the call center revealed that short service times were caused by agents that simply *hung-up* on customers. Agents used this disconnecting device to obtain extra rest-time. (We have since discovered that the phenomena of agents "abandoning" customers is not that scarce; it is, however, more often due to distorted incentive schemes, especially those that over-emphasize short average talk-time.) The problem was identified towards the end of 1999, after unreasonably many customers had complained about being disconnected: at least 10 months to unravel a problem that was "well hidden within the averages" that call center managers are typically using. Short service times are no longer prevalent in the November/December columns of Table 40, as well as in the histogram for Nov–Dec in Figure 17. (For some more details, see Subsection 8.2 that deals with the performance of individual agents.)

Tables 41–48 present summary statistics of service time (truncated at 1 hour of service) for the four main types: PS, NE, NW and IN, and for types PS and NE stratified additionally by priority 1 or 2. (NE and PS calls with priority 0 happen to be mostly calls that were transfered directly to an agent, hence their recorded priority is unlikely to be their true priority. Also, calls of type NW and IN are not stratified by priority since most have priority 0.)

Table 41: Service time, PS customers

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Ann
Mean	190	183	189	208	194	186	179	170	160	157	173	181	181
Med	122	116	119	132	122	120	116	113	108	105	115	123	117
SD	216	207	221	238	232	214	211	194	179	179	191	189	207
Tot	20551	18533	19704	17395	22365	22948	21854	21817	18677	21529	22945	23807	252125

Table 42: Service time, NW customers

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Ann
Mean	107	106	113	107	111	125	128	116	106	91	104	126	111
Med	61	60	65	58	60	66	74	68	62	59	67	78	64
SD	150	145	150	147	152	186	178	163	150	116	140	154	154
Tot	4054	5056	3462	2508	2908	3686	3199	3936	2946	3852	3280	2722	41609

Table 43: Service time, IN customers

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Ann
Mean	264	267	304	331	333	298	248	288	450	426	410	391	381
Med	166	159	160	204	192	163	121	119	243	243	219	220	196
SD	292	349	421	348	434	419	424	452	541	521	493	452	485
Tot	208	196	193	200	248	310	658	2179	2630	2485	2588	3111	15006

Table 44: Service time, NE customers

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Ann
Mean	321	275	273	341	301	257	242	237	231	242	260	280	269
Med	199	177	165	200	188	159	158	163	147	153	168	181	169
SD	376	301	329	406	357	332	290	255	290	289	294	318	320
Tot	1476	2516	2938	2432	3544	3404	3050	2988	2503	2719	3958	3722	35250

Table 45: Service time, PS customers, priority 1

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Ann
Mean	176	180	173	191	184	174	169	159	148	141	165	167	170
Med	111	114	114	124	116	113	107	104	99	94	109	114	110
SD	210	208	195	223	225	199	214	187	170	171	182	176	199
Tot	3245	3134	4389	3737	4374	3889	4903	4225	2558	2435	4447	5051	46387

Table 46: Service time, PS customers, priority 2

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Ann
Mean	200	193	201	223	213	200	194	186	170	180	193	199	198
Med	127	121	124	141	135	128	124	125	115	118	127	134	127
SD	228	216	242	256	246	232	223	211	189	207	211	203	225
Tot	6910	6524	9750	7947	9557	8493	9886	8743	5018	4084	8818	9518	95248

Table 47: Service time, NE customers, priority 1

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Ann
Mean	229	206	203	221	213	169	240	207	186	159	181	213	203
Med	137	126	124	121	129	97	149	138	103	111	126	142	127
SD	276	295	252	289	240	241	276	216	246	163	176	229	239
Tot	113	181	342	272	370	253	397	336	290	237	534	600	3925

Table 48: Service time, NE customers, priority 2

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Ann
Mean	299	268	281	350	305	265	228	234	232	255	269	281	274
Med	194	175	171	208	191	170	159	172	158	169	177	182	177
SD	326	296	334	410	356	317	254	244	278	299	294	327	319
Tot	648	1275	1924	1514	2200	1859	1713	1420	869	791	2129	1881	18223

We see that IN customers have the longest service times and NE customers are next in duration. NW have the shortest service time. All this is consistent with the nature of the IN, NE and NW calls. (An important implication is that the workload that IN customers impose on the system is more than their share in terms of percent of calls. This is an operationally important observation, and we return to it in Subsection 9.2.) A comparison between priorities 1 and 2 reveals that the latter (high priority) customers tend to have longer service times.

Figures 19–20 (p. 46,47) compare the service time distribution of the four types by estimating their densities, using kernels [30], and by looking at their survival functions. The densities for January–October do not recognize the short calls, as identified in Figure 17, because of the resolution (up to time 1200) and the bandwidth used. In Figure 20 we note a clear stochastic ordering between the types and priorities, which strengthens previously-discovered inequalities between mean service times. Thus, an IN customer is not only served longer than PS on average, but a PS service is more likely than IN to end at any given time t. Similar interpretations hold for other types and priorities.

Hazard rates are also informative for service times. (The most prevalent service times in Queueing Theory enjoy constant hazard rates, being exponential by assumption for a mere analytical tractability. As amply demonstrated in Figure 21, the exponentiality assumption is unjustified in our call center.) Hazard rate estimation is easier here than with waiting times since there is no censoring. As before, the empirical hazard rate at time t is the ratio of the number of calls ending at a time within $[t, t+1)$ after start, to the number of calls with service durations that exceed t. Here as before, this estimate becomes unstable as time increases since the remaining population diminishes. We used HEFT to "smooth" the hazard rates as a function of time (see Subsection 6.3 and [24]), and compared it with other smoothing methods. (Regression is also an alternative which we have not yet employed.) Using supersmoother (and other nonparametric methods) gave qualitatively the same pictures as HEFT. The bottom left plot (vertically) in Figure 21 (p. 48) shows the HEFT estimate of the hazard rates, superimposed on the empirical hazard rates. Note again the arches occurring at the right tails, which were already explained in Section 6.3. The other plots in Figure 21 show HEFT estimates of the hazard rates for different types and priorities. These pictures are consistent with the stochastic ordering that was observed in Figure 20. We do not fully trust the curve for type NW after 200 seconds due to the small size of the remaining population at this point.

In Figure 18 (p. 43) the HEFT estimate for the service time hazard rate was plotted for the first ten months of the year (left plot) and the last two months (right plot). The effect of the prevalent short service times during January-November is noticeable here.

Figure 18: Hazard rate for service time, Jan–Oct and Nov–Dec

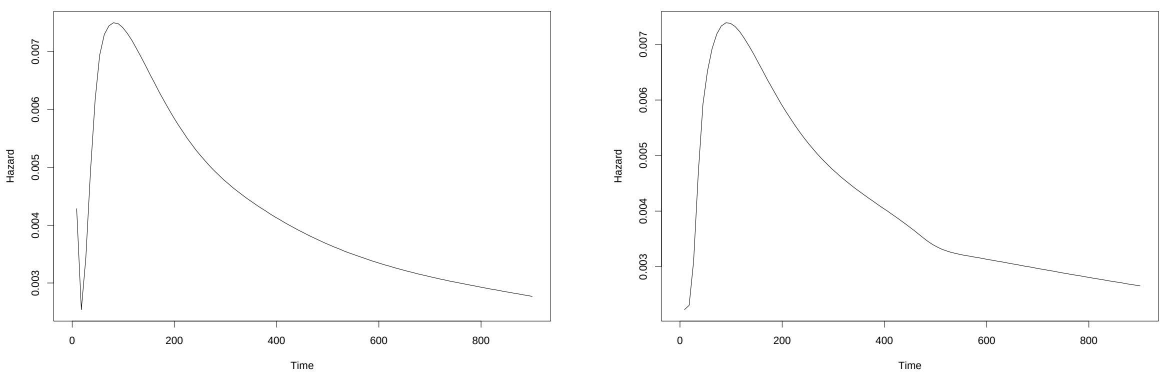


Table 49 summarizes the mean, SD and median service times in November and December for the different types using the tail formula. Note the similarity with the values for November and December in Tables 38 and 41–48. This is not surprising as there is no censoring for service times (differences are due to numeric integration).

Table 49: Means and medians of service time by types and priorities (Nov + Dec)

Service time	Mean	SD	Median
General	202	272	122
PS	178	206	119
NE	274	349	174
NW	115	176	72
IN	408	505	220
PS, 1	168	208	111
PS, 2	197	217	131
NE, 1	198	207	135
NE, 2	280	370	179

7.1 On service times and Queueing Theory

Most applications of queueing theory to call centers assume exponential service times as their default. The main reason is the lack of empirical evidence to the contrary, which leads one to favor convenience and apply models that are analytically tractable (for which there are readily available formulae.) And indeed, models with exponential service times are amenable to analysis, especially when combined with the assumption that arrival processes "are" Poisson processes (a rather natural one for call centers.) The prevalent Erlang-C is such a model – denoted formally by M/M/N, the first "M" stands for the assumption of Poisson arrivals, the second for exponential service times, and "N" is the number of agents.

In classical queueing formulae, the service time often affects performance measures through its squared-coefficient-of-variation $C^2 = E^2/\sigma^2$, where E is the average service time, and σ its standard deviation. For example, a useful approximation for the average waiting time in an M/G/N model (Poisson arrivals, General service times, N servers), is given by [15]

$$E[\text{Wait for M/G/N}] = E[\text{Wait for M/M/N}] \times \frac{(1 + C^2)}{2}.$$

Thus, average wait with general service times is multiplied by a factor of $(1 + C^2)/2$ relative to the wait under exponential service times. For example, if service times are in fact exponential then $C^2 = 1$ and the factor is 1, as should be; deterministic service times *halve* the average wait (think of changing human servers to robots); and finally, our empirical service time from Table 38 amplifies it by $[1 + (304/190)^2]/2 = 1.78$. (With Table 39, the factor is 1.31 – decreasing stochastic variability reduces wait.)

In the approximation above, and many of its "relatives", service times manifest themselves merely through their means and standard deviation. Consequently, for practical purposes, if means and standard deviations are close to each other, then one can assume exponentiality of service time. However, for large call centers with high levels of agents' utilization, simulation studies (by colleagues at Bell Labs) indicate that the *whole* distribution of service time may become significant. As clear from Figure 17, the distribution of service times is *not* exponential, and significantly so. To wit, its C^2 is larger than unity (see also Table 49), and its histogram/density does not have the exponential shape (compare with Figure 10).

We have been thus lead to the important issue of fitting a parametric statistical model to service times. Research by our colleagues at Wharton has revealed a *remarkable* fit to the *log-normal* distribution, not only for the overall service time, but also when restricted to service types, individual agents, etc. It follows that the natural log ($\log_e$) of service time is *normally* distributed. The implications of these findings are presently being explored. One example is the analysis of covariates that affect service time, by simply applying standard regression techniques to log(service time).

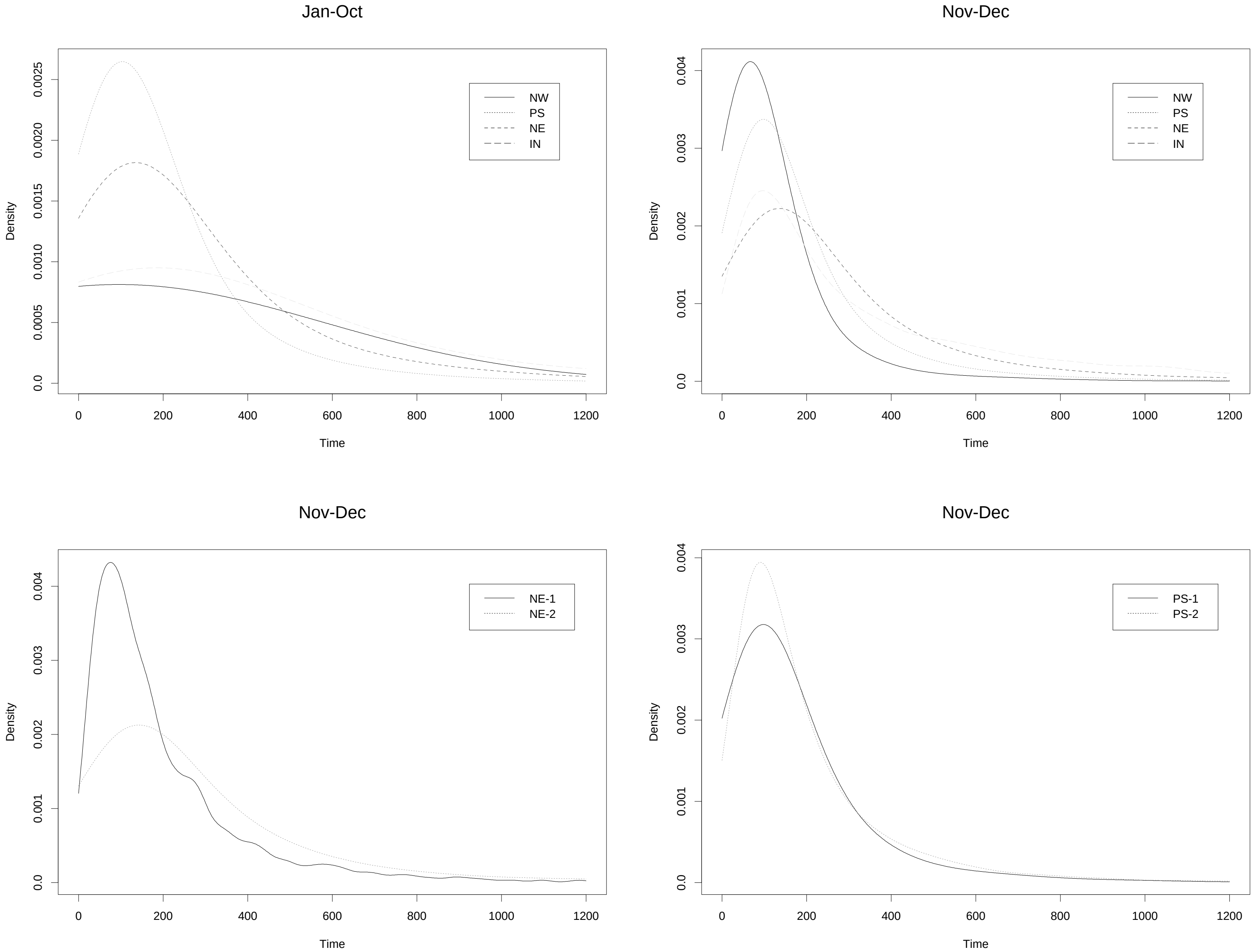


Figure 19: Densities of service times, by types and priorities

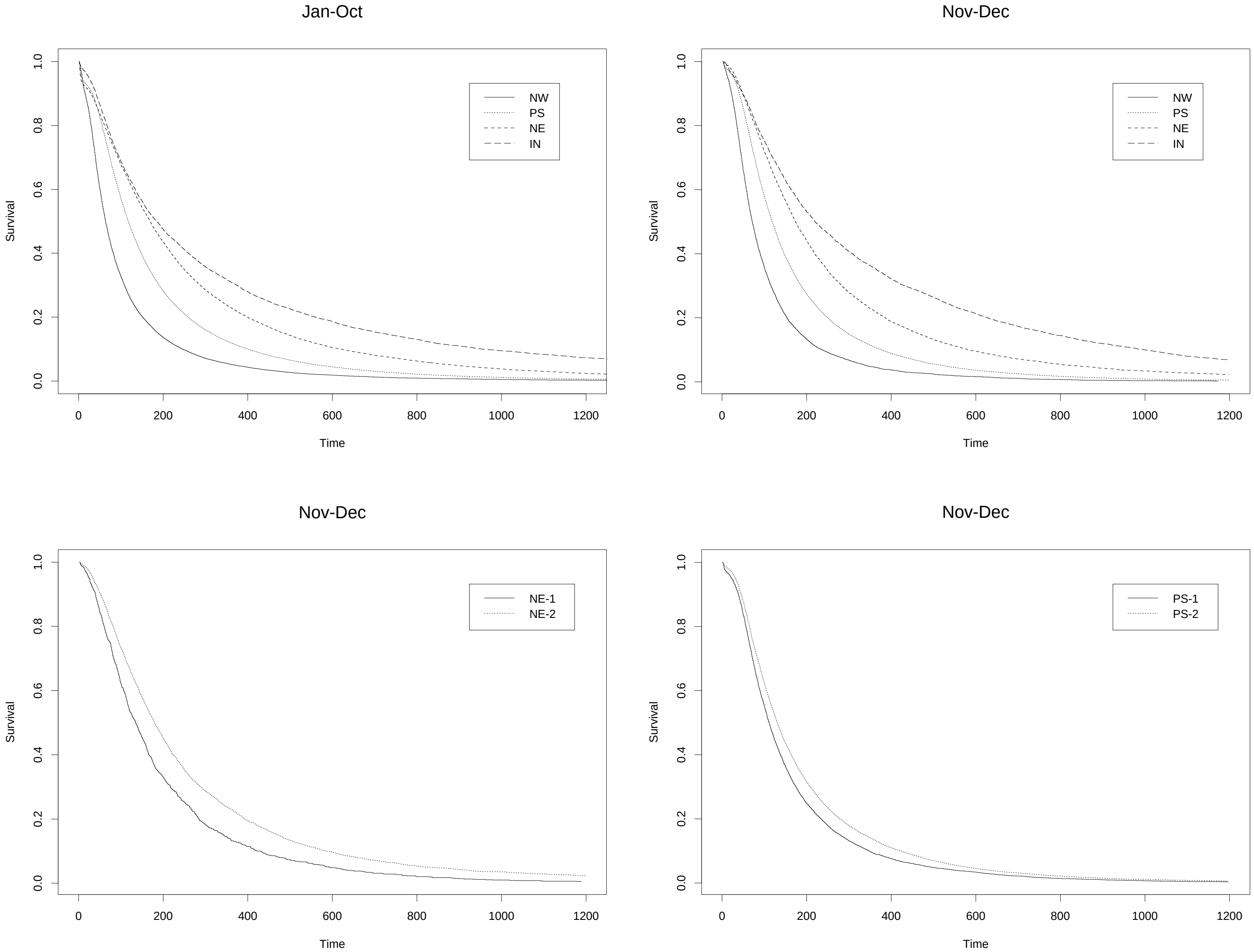


Figure 20: Survival function of service time, by types and priorities

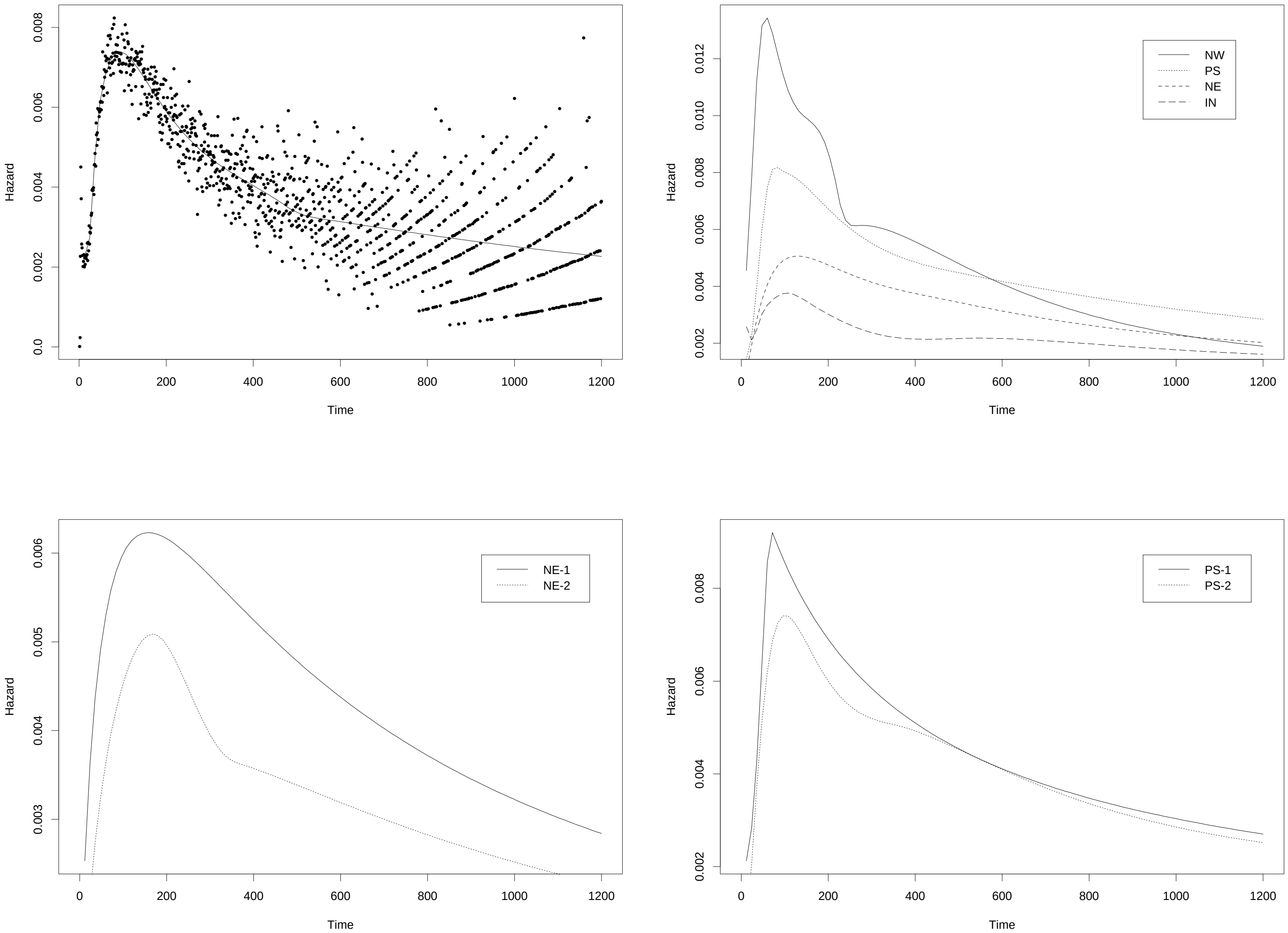


Figure 21: Hazard rate for service time, by types and priorities (Nov + Dec)

8 Individual behavior of customers and agents

An obvious significant advantage of transactional data, such as ours, is that it enables the analysis of individual customers and agents. In Section 8.1 we first look at the number of calls that individual customers perform over a month. We then focus on two very frequent callers. In Section 8.2 we describe the number of customers that various agents handle over a month, and then focus on one examplary agent.

8.1 Analysis of individuals customers

Most customers call several (sometimes many) times over the year. Tables 50–52 summarize the distribution of the number of calls made by a single customer, for each month and annually. The first row in the table, labeled "# ID's", presents the number of distinct customers calling each month, and during the year. The next 5 rows give summary statistics: mean, SD and maximum. The distribution is skewed to the right, and its 99*th* percentile appear in the row labeled 99% (i.e. 1% of the customers had a larger number of calls than the number in the corresponding cell, during a given month). Table 50 covers all calls, while Tables 51–52 are for PS and NE customers.

Important remark: Recall that only about 47% of the calls were identified. (For about half of the calls, the average customer was lucky enough to be served immediately upon leaving the VRU, in which case no ID was registered.) A reasonable assumption is that these provide a representative enough of a sample, so the actual frequency of calls by an individual customer is about twice of what was reported above.

Some customers are using the system only part of the year. We deduce this when looking at the #ID's row in Table 50: the number of distinct customers calling each month ranges from about 3000 to 5000. If the same group was calling every month, then the number of distinct customers calling during the year should have been about 5000, but it is closer to 13,000. The same applies to Tables 51–52. This also explains why the mean 16.3 of the column "Total" is much smaller then $3.5 \times 12 = 42$, where 3.5 is the mean of row "Mean" over the year: thus, an average customer uses the system about 3.5 times per month, but only during some months and not during others. One explanation could be that new customers were joining while others were leaving the service of the call center. The in- and out-flows seem rather balanced, with a slight increase towards inflow at year-end. Naturally, the joining-leaving process of customers is very important to track, and managers of call centers ought to monitor it closely and continuously.

Note the difference between the frequency of calls by PS and NE customers: the formers call about 3.5 times per month, while the latters calls about 18.5 times per month. However, there are only 305 distinct customers that use the NE services of our call center.

Table 50: Distribution of number of calls by an individual customer

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Total
# ID's	3976	4081	5089	4657	4906	4444	4938	4743	3463	3155	4266	4688	12902
Mean	3.4	3.4	4.6	3.9	4.4	4	4.4	4	3	2.7	4.7	4.7	16.3
Median	2	2	2	2	2	2	2	2	1	1	2	2	5
SD	8	8	12	9	12	11	11	10	7	6	13	12	64
99%	39	35	55	42	60	53	59	46	34	33	68	69	229
Max	151	170	238	171	284	219	235	204	137	103	279	189	1996
Total	13424	13786	23365	18209	21716	17888	21847	19048	10300	8543	19910	21860	209896

Table 51: Distribution of number of calls by an individual PS customer

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Total
# ID's	3902	3953	4971	4527	4782	4323	4815	4623	3349	3037	4145	4568	12701
Mean	3	3	4	3	4	3	4	4	3	2	4	4	14
Median	1	1	2	2	2	1	2	2	1	1	2	2	4
SD	7	6	9	8	10	9	10	9	6	5	11	10	51
99%	31	24	39	36	39	37	45	40	27	24	53	52	163
Max	125	170	238	171	283	219	235	204	137	103	279	179	1996
Total	11730	11231	19010	15025	17630	14679	18467	16531	8975	7256	16207	18288	174849

Table 52: Distribution of number of calls by an individual NE customer

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Total
# ID's	77	98	117	121	134	119	123	105	93	100	109	149	305
Mean	11	17	24	17	22	20	20	19	13	11	28	19	83
Median	5	7	9	8	10	7	10	11	7	4	9	5	5
SD	18	25	33	23	31	26	27	24	16	15	38	29	188
99%	79	85	133	105	132	113	105	114	66	52	163	128	929
Max	136	170	182	141	166	156	150	119	71	75	204	153	1471
Total	846	1643	2759	2062	2972	2402	2491	2000	1246	1086	3049	2830	25386

8.1.1 The most obsessive callers

As seen above, the average customer calls about 16 times over the year. There is however a customer who made (at least) 1471 calls of type NE during the year, and another with (at least) 1996 calls of type PS. (The latter translates to about a call per hour.) We now look closely at the call history of these two "obsessive callers". (Recall that we are analyzing only those calls during which the customers spent a positive time waiting.)

The caller with 1471 calls of type NE reached an agent in 95% of his calls, and abandoned the rest. His average queueing time (over all calls) is 108 seconds. His average service time is 586 seconds. (Thus, on average he occupied an agent for about 50 minutes per day.) Naturally, the call center staff knows this customer. We were told that his long service times are for consultation on his stock-trading.

The other customer, with 1996 calls of type PS during the year, reached an agent in 92% of his calls. His average queueing time was 117 seconds, with an average service time of (curiously also) 117 seconds. We were puzzled by the volume of PS (standard) calls performed by this individual. But the call center staff easily resolved it for us: this customer was actually performing stock-trading (i.e. NE type services). His calls were relatively short since he had been gathering all needed information from the Internet. Then why PS calls? The answer turns out simple: the customer was calling long-distance to the bank, and NE calls are caller-paid while PS calls are *toll-free*. Thus, the customer was calling PS calls in order to perform NE activities, saving significant indirect transaction costs when doing so. This raised the general issue of the validity of our segregation into call types. The call center staff assured us of the rarity of the situation, and that the vast majority of customers are indeed getting the service that they are dialing for.

8.2 Analysis of individual agents

Agents are the most important resource of a call center. They are also a challenging resource to manage: Salaries constitute between 60% to 70% of operating expenses, training is non-trivial with some skills being scarce, yet attrition rates are typically high. As will now be demonstrated, our data allow the analysis of the operational dimensions of agents performance. Here we emphasize service times. Such an analysis ought to supplement more prevalent modes, for example on-line and off-line monitoring, and customers/agents surveys.

An important aspect of being able to study the service time of an individual agent is the ability to follow *learning curves*: one expects a new agent to provide slower services than an experienced one. Learning curves affect, among other things, economically significant staffing decisions. During 1999, our call center was stable in terms of its agents: the overall number varied little, agents typically worked full time and very few joined during the middle of the year. (Some agents did not answer calls for a month or two over the year, plausibly due to a vacation or a temporary assignment elsewhere at the bank.) Thus, our data-set is not very promising for learning-curve analysis, with an exception being some of the agents who provided IN type of service (see below). (Future studies of large-scale call centers ought to pursue this direction.)

But even with similar experiences, call center agents still constitute a heterogeneous population, widely varying in capabilities and skills. We are thus able to trace extreme performers (both at the low-end, and stars at the high-end), which is important for understanding performance limits, then setting goals and designing incentives. The analysis of individual performance increases in importance with the growing practice of skills-based-routing [12]: the Automatic Call Distributor (ACD) routes customers to agents by matching, as best as possible, customers' demands to agents' skills. Understanding and tracking these skills becomes therefore important.

Phone calls with positive service time and server status NO_SERVER were excluded from our analysis. Most agents provided all types of services (including IN). Starting in August, a special new group of agents was assigned exclusively to IN services. Agents of this group have a prefix 'Z1' or 'Z2' added to their name. Note that the other agents still provide occasional IN services. (With regard to the learning curve, we see that for some agents, the number of customers being served increased from one month to the next, which suggest that they indeed 'learn'. However, we do not pursue this further here).

Table 53 shows the number of calls that agents handled each month. Included are agents who worked for at least 8 months during the year (with MEIR being an exception, since he seems to be the only one to have joined the center at the middle of the year). For the Internet group, we include agents who worked for at least 4 months.

Table 53: Number of calls handled by an agent

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
AVI	0	0	0	1117	2208	2019	2789	2710	1417	2026	2523	2395
AVNI	1493	1736	642	539	1786	2219	2092	2392	1156	1888	1988	2136
BASCH	999	1164	1708	1155	982	906	858	2185	1973	1055	1326	1242
BENSION	1283	1135	0	1053	1108	1016	1682	1298	1076	1303	1546	1176
DARMON	309	515	633	519	577	436	309	370	297	194	425	128
DORIT	696	1047	0	811	546	862	750	2228	1319	1384	1640	1605
ELI	387	508	777	447	560	436	395	458	416	363	502	352
GELBER	333	143	510	427	859	281	386	332	67	179	165	269
GILI	668	614	1155	803	1108	974	418	0	355	456	412	298
KAZAV	1995	1693	1240	1451	1731	2251	1737	1168	729	1570	1047	2038
MEIR	0	0	0	0	0	0	127	344	318	280	406	454
MORIAH	1360	1223	1591	1351	1866	1980	2416	2152	1526	1940	1793	515
PINHAS	79	40	359	244	31	311	422	241	143	105	51	63
ROTH	0	0	397	1292	1928	1967	1831	1749	1625	1914	1458	1038
SHARON	1985	1674	2780	1938	2563	2657	2537	2875	1803	1935	2532	2140
STEREN	0	1043	2294	1516	2163	2231	1423	2455	1672	709	2375	2568
TOVA	1923	1679	1562	1059	1464	1389	1890	1811	1361	1971	941	0
VICKY	895	0	0	0	1006	1378	1415	1674	1472	1582	1641	1990
YIFAT	1312	1901	1745	1305	1464	1076	780	90	1137	1315	0	0
YITZ	1771	1791	1402	1203	1355	1367	1009	69	705	1743	2420	2353
ZOHARI	891	1144	1398	1148	1479	1450	980	1494	1423	1359	1504	1094
Z2ARIE	0	0	0	0	0	0	0	56	225	315	432	534
Z2ELINOR	0	0	0	0	0	0	0	45	352	288	222	310
Z2EYAL	0	0	0	0	0	0	0	95	331	428	579	618
Z2IFAT	0	0	0	0	0	0	0	94	260	314	215	0
Z2LIOR	0	0	0	0	0	0	0	84	250	136	126	138
Z2NIRIT	0	0	0	0	0	0	0	116	327	474	387	545
Z2OFERZ	0	0	0	0	0	0	0	71	311	260	242	334
Z2SPIEGEL	0	0	0	0	0	0	0	71	311	260	153	322

Table 54: Number of calls with short service time

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
MORIAH	233	230	356	290	614	695	865	597	490	455	4	1
AVI	0	0	0	47	111	144	295	221	121	76	35	26
AVNI	11	13	4	5	6	25	16	18	4	8	8	11
DARMON	2	11	8	9	10	7	1	0	1	1	0	0
ELI	9	7	10	12	22	18	15	4	8	3	6	5
KAZAV	57	40	48	44	48	63	40	27	15	18	4	6
MEIR	0	0	0	0	0	0	1	8	3	1	2	1
PINHAS	3	0	58	25	4	14	11	6	8	1	0	0
ROTH	0	0	10	10	36	21	43	25	32	31	3	6
SHARON	58	49	86	52	67	78	66	63	38	23	43	49
TOVA	52	163	269	132	231	193	100	109	207	190	6	0
ZOHARI	4	8	12	22	17	20	9	14	5	7	10	7

Recall the phenomenon of agents "abandoning" customers, which was discusses in Section 7. (See also Figure 17.) The problem of short services (under 10 seconds) was, by far, most severe for MORIAH. In Table 54 we show the number of calls with short service time, for MORIAH and some of the other agents. (Note MORIAH's dramatic "improvement" at the end of the year, when the phenomenon of disconnecting calls was unraveled.)

Tables 55–59 display operational characteristics for five of the agents: their service-mix and the overall mean, SD and median of their service time. The rows labeled 'PS', 'NE' etc. show the percent of calls from the given type. The last row, labeled %, shows the percent of short calls. We see that the service-mix varies among agents; for some, it is pretty close to the distribution in Table 10 (p. 19), and for others it is rather different. It seems that ELI is the local expert on NE calls (note the high faction in Table 57). Note again the high percentage of short calls for MORIAH (Table 58), and the "improvement" in November–December.

Table 55: Characteristics of AVNI

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
PS	74	68	70	81	82	76	76	72	74	80	80	77
NE	5	5	2	1	1	3	2	9	11	7	3	4
NW	18	23	21	12	12	18	19	19	15	12	13	16
TT	2	4	5	5	4	2	2	0	0	1	4	4
PE	0	0	1	0	1	1	0	1	0	0	0	0
Mean	282	257	299	268	264	200	228	209	215	209	245	221
SD	291	251	312	269	258	191	226	197	227	252	292	214
Median	188	177	198	179	183	138	152	147	154	145	160	152
%	0.7	0.8	0.6	0.9	0.3	1.1	0.8	0.8	0.3	0.4	0.4	0.5

Table 56: Characteristics of DARMON

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
PS	83	75	73	77	72	77	75	84	85	85	82	73
NE	9	17	16	15	18	15	20	9	13	13	16	23
NW	1	3	3	1	1	1	1	2	0	1	0	1
IN	0	0	0	1	0	0	2	1	2	1	0	0
TT	6	5	7	6	9	7	2	4	0	1	1	3
Mean	286	250	243	257	282	274	273	285	273	295	270	300
SD	362	284	273	266	356	324	279	305	319	367	520	339
Median	173	153	155	179	172	160	180	182	165	155	176	198
%	0.6	2.1	1.3	1.7	1.7	1.6	0.3	0	0.3	0.5	0	0

Table 57: Characteristics of ELI

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
PS	24	10	23	13	9	2	4	3	17	10	5	9
NE	71	89	74	82	91	98	94	96	83	90	95	91
NW	1	1	0	1	0	0	2	0	0	0	0	0
TT	4	0	2	4	0	1	0	0	0	0	0	0
Mean	445	394	382	482	487	554	431	385	438	447	440	497
SD	479	424	417	557	578	1085	491	371	595	459	539	842
Median	254	253	248	280	293	328	266	277	250	289	292	281
%	2.3	1.4	1.3	2.7	3.9	4.1	3.8	0.9	1.9	0.8	1.2	1.4

Table 58: Characteristics of MORIAH

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
PS	72	59	68	69	68	62	70	67	65	62	51	51
NE	7	16	15	14	16	18	12	11	15	13	13	11
NW	19	24	16	16	15	20	16	20	18	23	33	37
PE	2	0	1	1	1	1	1	1	1	1	1	0
Mean	203	193	189	215	191	163	140	151	142	127	154	190
SD	247	242	239	314	291	254	221	223	224	181	163	209
Median	125	110	111	109	75	71	65	83	75	77	102	111
%	17.1	18.8	22.4	21.5	32.9	35.1	35.8	27.7	32.1	23.5	0.2	0.2

Table 59: Characteristics of ZOHARI

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
PS	67	45	70	55	65	74	61	63	71	75	68	61
NE	31	46	24	37	33	25	38	28	25	24	30	38
NW	2	5	4	4	2	0	0	8	4	2	2	1
TT	0	3	2	4	0	0	1	0	0	0	1	0
Mean	258	252	267	292	285	274	274	236	217	233	269	337
SD	327	268	295	305	303	312	280	256	203	225	290	425
Median	155	162	175	199	188	186	189	160	154	173	180	211
%	0.4	0.7	0.9	1.9	1.1	1.4	0.9	0.9	0.4	0.5	0.7	0.6

In Table 60 we summarize the characteristics of MEIR, who apparently started working in July. We see that MEIR's mean service time does not go down with time. (Actually, the type-mix of services changed somewhat over July-December, so mean service times ought to be compared per type of service, which we have not done.)

Table 60: Characteristics of MEIR

	Jul	Aug	Sep	Oct	Nov	Dec
PS	59	70	75	70	68	69
NE	14	11	14	14	20	16
NW	19	19	11	16	12	15
IN	8	0	0	0	0	0
TT	0	0	1	0	0	0
Mean	207	186	179	183	202	212
SD	225	173	188	229	236	248
Median	130	135	124	116	131	128
%	0.8	2.3	0.9	0.4	0.5	0.2

Finally Table 61 shows the mean, SD and median service time of four agents who specialized in Internet services. The fraction of calls with short service time is very low (under 0.5%), hence it was excluded from the table.

Table 61: Characteristics of Internet agents

		Aug	Sep	Oct	Nov	Dec
	Mean	590	635	531	482	438
Z2ARIE	SD	785	787	542	590	496
	Median	290	354	361	277	264
	Mean	860	524	425	372	404
Z2ELINOR	SD	892	608	480	426	529
	Median	428	285	249	213	215
	Mean	406	455	417	366	342
Z2EYAL	SD	481	531	630	494	411
	Median	243	253	200	183	191
	Mean	381	418	435	375	384
Z2NIRIT	SD	568	560	550	453	442
	Median	131	218	250	207	214

8.2.1 An examplary agent: ZOHARI

In this subsection, we analyze in more details the service times of ZOHARI. In Tables 53, 54 and 59 we saw the total number of calls that ZOHARI handled during the year, and the fraction of calls with durations shorter than 10 seconds. We also saw the service types that ZOHARI covered, then the overall mean, SD and median of her service times. A similar analysis to the one presented now, can be easily carried out for any other agent. We chose ZOHARI since she was working all year, with a relatively large volume of calls, and only few short calls. Our analysis covers all calls with positive service time, and it excludes a single PS call with service time of 8313 seconds.

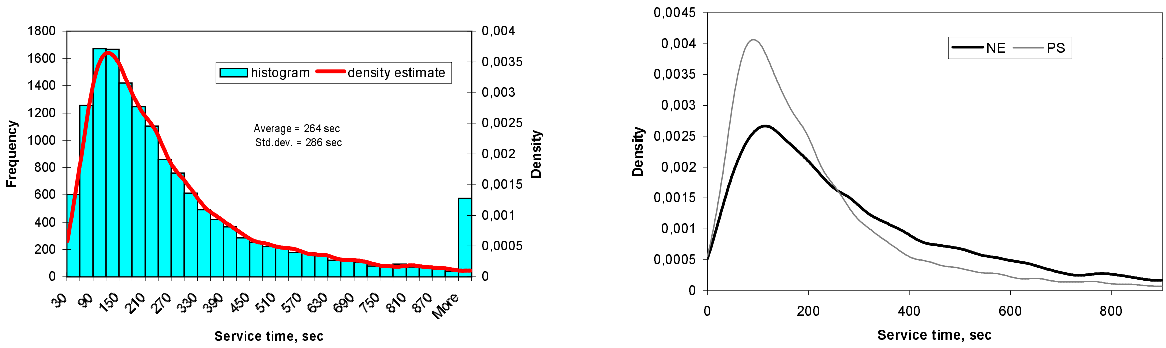
Table 62 summarizes ZOHARI's mean, SD and median service time during the year, overall and for the different service types. The column 'Total' shows the total number of calls included in the analysis. (There was also one call of the outlier type PE, which is not shown in the table.)

Table 62: Some more characteristics of ZOHARI

	Total	Mean	SD	Median
Overall	15363	265	293	176
PS	10029	234	251	161
NE	4761	347	354	235
NW	418	123	138	77
TT	154	137	175	71

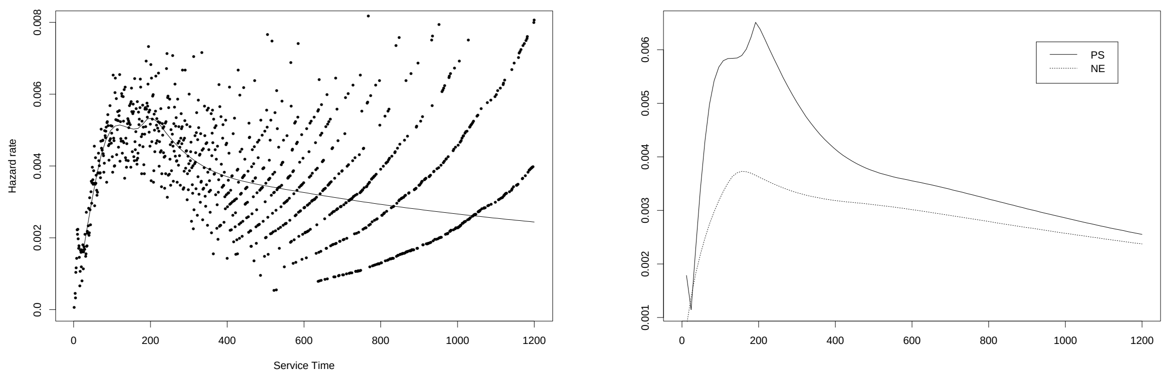
Figure 22 shows the distribution of ZOHARI's service time: the left plot is the histogram of all of ZOHARI's service times, superimposed with a kernel density estimate, and the right plot shows the density estimate of ZOHARI's service time for types PS and NE.

Figure 22: ZOHARI's service time distribution



Next we look at the hazard rate for ZOHARI's service time. The hazard rates were smoothed using HEFT [24]. Figure 23 shows the HEFT estimate, superimposed on the empirical hazard rates, and the HEFT estimate for types PS and NE. The shapes of both the density and the hazard rates are similar to those observed for the overall agent population; see Section 7.

Figure 23: Hazard rate for ZOHARI's service time



9 Call dynamics during an individual day

In Figures 2–5 (p. 15–17), we presented four different levels for describing a call center. Here we re-examine the operational level, and the present section can be broadly thought of as refinements to Figure 4.

From a practical point of view, this section is oriented towards the operations manager of a call center, who plans and controls hourly and daily performance. The figures presented below ought to constitute a manager's daily input, for both planning and feedback analysis. From a scientific point of view, the section caters to queueing theorists, whose typical steady-state models are used to fit hourly operations of call centers. (For example, the Erlang-C (M/M/N) model, conceived already in 1917, has been the most common model to support hourly staffing decisions; these are then translated first into shifts and ultimately into individual agents' assignment; see [6], written already in 1976 but still relevant, for more details.) It is now increasingly recognized that the staffing reality of today's call centers is far more complex than that allowed by Erlang-C, and research that accommodates at least part of this complexity is now appearing.

Figure 4 clearly manifests that the arrival process of calls is not homogeneous in time over the day: some periods are more heavily loaded than others. This is the reason behind the practice of fitting queueing models to a single hour – most such models assume (for tractability) steady-state conditions. Such conditions could perhaps apply over a relatively short (hence steady) period, but even then only if traffic volume is high enough (hence steady state is reached fast enough).

We start the section by refining Figure 4, thus remaining at the deterministic "fluid-like" levels of description, in terms of flow-rates. For example, the analog of Figure 4, according to service types, reveals differing daily behavior across types. (Deeper statistical analysis, aimed at identifying distributional characteristics, is left for future research. For example, does the arrival process fit a time-inhomogeneous Poisson process? And if not, what does it fit? See Subsection 4.3 for the assumptions that justify a Poisson model.)

The arrival rate is only one important characteristic of daily call dynamics. Since a call goes through several service phases (VRU, Queue and Service), it is of interest to examine the interrelations between these phases, and their relation to performance measures. Consider, for example, the dependence on arrivals of the fraction who abandon, or of the average waiting time. For an analysis of this dependence, one must combine the arrivals with the service they seek. (Indeed, few long IN services could impose the same load as many short PS services.) This gives rise to the concept of average *workload*, used below as the product of the average arrival rate (during a given time interval) with the average service time (over that same interval). Other processes of interest are *counting* processes that record the (average) number of customers at the VRU, at the queue, or being served, as a function of time.

In the first part of this section (Subsections 9.1 and 9.2), we consider characteristics of daily dynamics, for a typical day (in November). We saw in Figure 6 the periodicity in the number of calls arriving over the week (ignoring holidays). The figure also reveals a few days that are unusual, as far as the number of arriving calls is concerned. Specifically, 2000 calls arrive on a typical weekday, but note the jump on May 23*rd* (3064 calls) and on July 4*th* (2589 calls). In Subsection 9.3 we examine more closely what happened during these two unpredictable days.

9.1 The November data

We consider November weekdays (22 days), since the number of calls during this month was large, there were no holidays and the problem of short service times had been corrected. We consider only calls which reached the center during working hours (17 hours a day, from 7:00am to 12:00am on Sunday to Thursday), whose outcome was AGENT or HANG, and which were not VRU abandonment (i.e. if the outcome is HANG then the call has a positive time in queue). There were 36,409 such calls (out of the 41,019 calls in November) and their distribution according to the four main types of service is: PS – 65.9% of the calls; NW – 12.2%; NE – 11.6% and IN – 7.7%.

9.2 A typical day in November

Our findings on daily characteristics are all presented graphically. For each characteristic, two figures are displayed: the first shows the overall level of the characteristics, compared with its level for type PS. The two are typically close, as the majority of the calls are PS. If there are differences, they are attributed to some behavior of the other types. To this end, the second figure considers the same characteristic, but for types IN, NE and NW (and when found useful, PS is included as well.)

For the characteristics below, we averaged each of the 17 hours over the 22 weekdays. The reason for considering hourly averages, and not other time intervals, is that two-hour intervals, for example, exhibit too much predictable variability over the two hours (they are not homogeneous); while half-hour intervals have too much stochastic variability across the 22 days (they are too noisy).

In Figure 24, we consider the average arrival rate per hour (i.e. averaging over the 22 weekdays, we count the number of calls arriving to the center during each of the 17 working hours). The graph for all calls is the one in Figure 4. Arrival rates according to types show clear peaks for type NE at the opening and closing times of the stock market, and a sharp decrease after the second peak; IN arrival rates are rather stable during the day, with two moderate peaks: around 18:00h (maybe customers arriving home after work) and around 22:00h (reduced cost of phone-connection to the Internet). Type NW has a peak around 16:00h, which is unclear to us. Note that PS arrival rates and the overall arrival rates have similar daily pattern. However, the differences are larger at hours with noticeable peaks for the other types (and in particular, when NE peaks).

Figure 24: Average arrival rate per hour (Nov, weekdays)

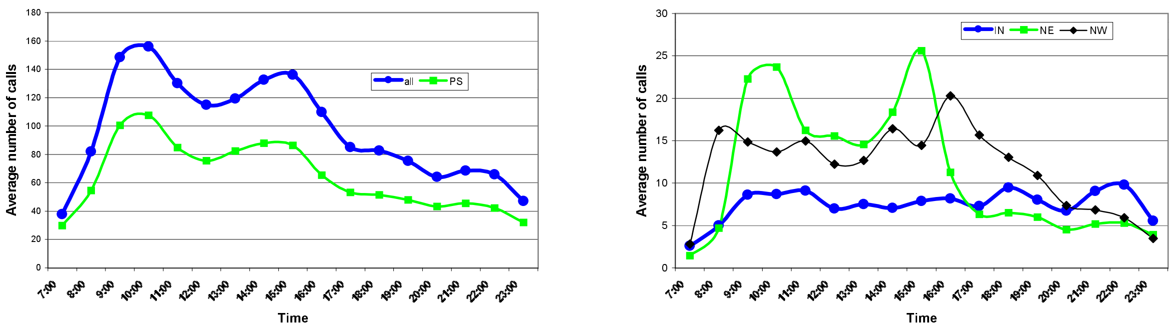
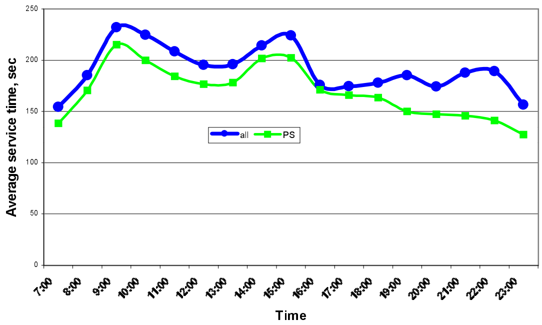
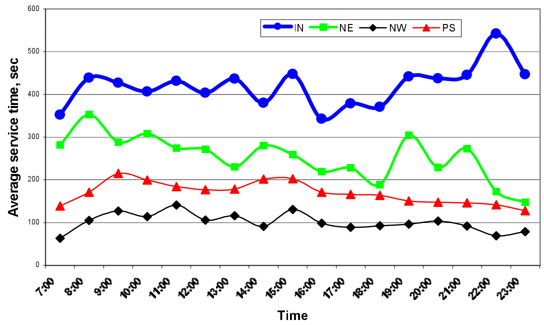


Figure 25 shows the average service time, calculated for each hour over the day, for all calls and segregated to types. Note that at all hours, the mean service time for type NW is smaller than PS, which is smaller than NE, which is yet smaller than IN. This was already noticed at the monthly level, in Tables 41–44 (p. 41–42), and in the stochastic ordering of the survival curves in Figure 20 (p. 47). For types NE and NW, the average service time seems rather stable over the day. The average service time for type IN seems to increase in the evening. Comparing the average service time for type PS with the overall, shows that the two are close to each other, with the largest differences at the peaks (which are precisely the busy hours for NE callers, who have relatively long service times – hence the increase in the overall average). Larger differences arise also during the late evening, with a similar explanation (but applied to type IN). On the other hand, we notice that the PS and overall averages are very close around 16:00h, where there is an increase of NW calls whose durations tend to be shorter (hence they pull down the overall average).

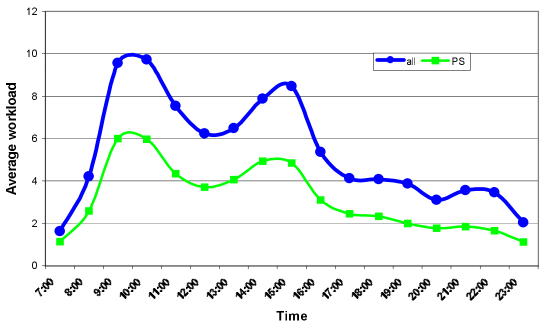
Figure 25: Average service time per hour (Nov, weekdays)

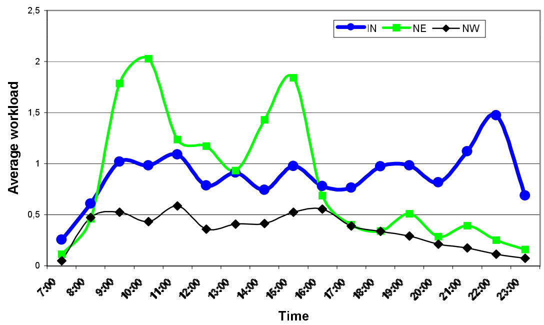




System congestion is a function of not only the arrival process, but also of the service times that these arrivals seek. We define the (average) *workload* to be the product of the average arrival rate, at a given hour, with the average service time, over that hour. Figure 26 shows this average workload. (Note: we multiplied the two averages for each hour. Alternatively, one could take the 22 products of arrivals by service time, at each of the 22 weekdays, and then average them.) Comparing, for example, IN and NW in Figures 24 and 26 reveals that, although IN customers call less, the IN share in workload is larger than that of NW callers. In particular, notice the workload for IN at evening time. The workload of NE callers is largest in the morning and before closing of the stock market.

Figure 26: Average workload per hour (Nov, weekdays)





How does workload affect the probability of delay (wait > 0)? We estimate this probability by calculating the total number of callers that waited a positive time at the queue, then dividing it by the total number of customers that called the center. The "total" in both cases is over 22 days, calculated for each hour separately. Figure 27 shows these probabilities, overall and according to types. (Recall that we have excluded from our analysis the customers who abandoned from the VRU.)

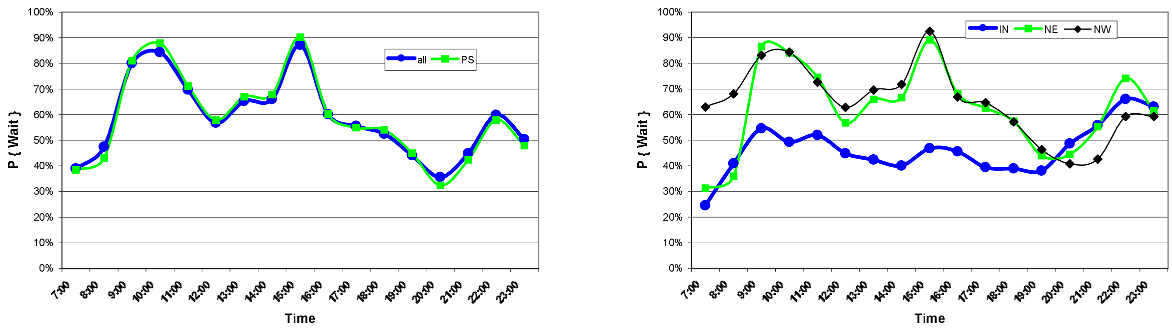
The probability of waiting increases around 9:00am, 15:00h and 22:00h, and it is almost identical to the delay probability for type PS only. Waiting probabilities are also very close for types NE and NW (after 9:00am), and both are close to type PS and the overall probability. There is a simple explanation for the similarity in delay probabilities across PS, NE and NW: these three types are close to be sharing the same queue (i.e. they are served by the same group of agents, yet some of the agents are more likely than others to serve specific type – see Table 57), according to a FCFS service discipline (after accounting for high priorities; see item 4. in Section 2). As for the overall probability, it includes IN callers as well, but there are relatively few of them, hence they have merely a small effect on the overall delay probability. Delay probabilities for type IN are different since IN customers are mostly served by a dedicated group of agents (see Section 8.2).

The reason behind the increased delay probability at 9:00am and 15:00h, for types PS, NE and NW, is the increased workload at those times (Figure 26). Our assessment for the peak at 22:00h is reduced staffing levels at late hours. (This finds some support later.)

Note the increased delay probability for type IN over the evening. During the middle of the day it is lower since, as already indicated, they are served by a dedicated group of agents, and they tend to call less during mid-day.

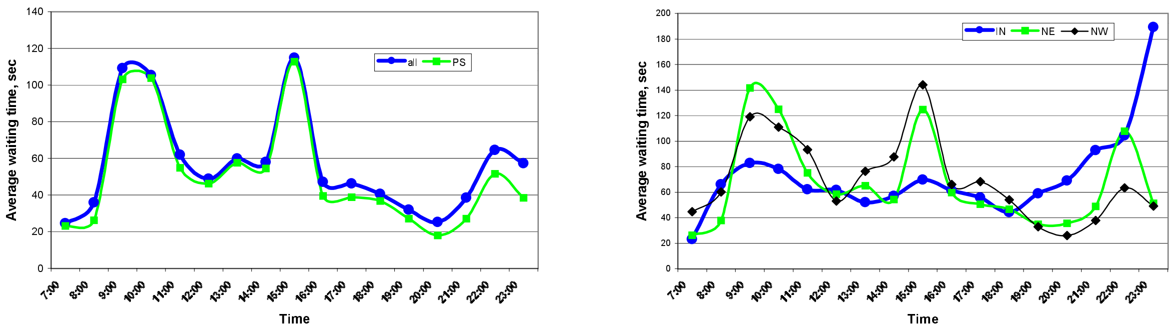
Finally, note that according to the priority protocol (item 4. in Section 2), the priority of customers affects their time in queue, but not their probability of being delayed in the queue.

Figure 27: Probability of waiting per hour (Nov, weekdays)



Another common measure of congestion is the average queueing time (which includes zero wait – the ideal from the customer's point of view, but very costly to the call center). In call center parlance, is it often referred to as ASA = Average Speed of Answer. One expects high correlation between average waiting time and the probability of waiting, as well as between the average waiting time and workload. Figure 28 shows the average waiting time in a given hour (averaged over the month), for all calls and according to type. As expected, the same pattern emerges for types PS, NE and NW. We do not expect the figures to be as close as in Figure 27, due to a combination of two reasons: the priority of customers affects their waiting time (NE are mostly priority 2, PS are mostly priority 2 but not as much as type NE, and NW callers have priority 0); and customers of different types behave differently, as demonstrated in Figure 15 (p. 37). Note that the average waiting times for types PS, NE and NW have the same pattern as the probability of waiting: all increase in the morning, before closing of the stock market, and during the evening. The difference between the overall waiting time and PS waiting time, during the evening, can be explained by the increased waiting time for type IN.

Figure 28: Average queueing time per hour (Nov, weekdays)

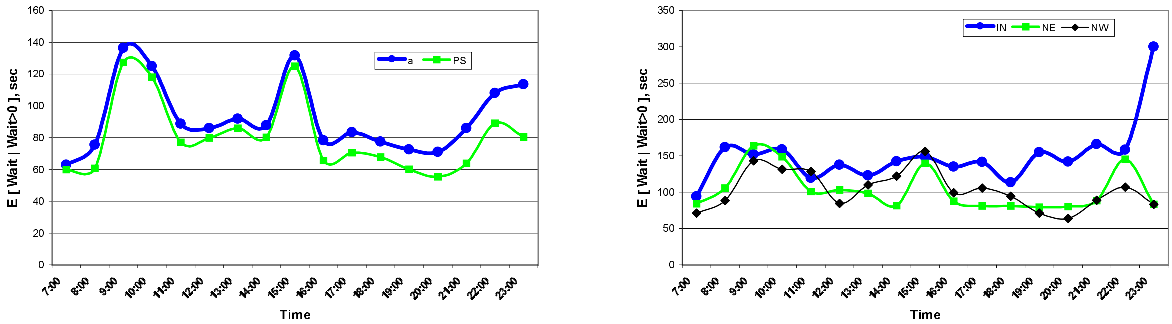


In Figure 29, we consider the average waiting time, here only for those customers who actually waited. We expect the picture to have the pattern of Figure 28, but with values that are larger. (Having omitted all customers with zero waiting time increases the average waiting time). The "mathematical" reason for the increased average conditional waiting time for type IN is:

$$E(\text{Wait} \mid \text{Wait} > 0) = \frac{E(\text{Wait})}{P(\text{Wait} > 0)},$$

and $P(\text{Wait} > 0)$ is low for type IN in mid day. Put in words, the explanation is as follows: IN customers enjoy their own group of agents, and during mid-day they call less, hence they are less likely to wait in the queue (Figure 27); which means that more of them have zero waiting time, thus reducing their average waiting time (Figure 28). On the other hand, when IN customers need to wait (for example at evening times), then they wait more time than in midday, since their average service time is longer than that of the other types.

Figure 29: Average waiting time per hour among customers who waited (Nov, weekdays)



One expects to see also high correlation between the probability of waiting and that of abandoning. The latter is shown in Figure 30. As can be expected, in light of previous figures and discussion, the probability of abandonment increases when customers must wait more (compare Figure 30 with 28). Note that NW customers have larger probability of abandonment than PS customer who, in turn, have larger probability of abandonment than NE customers. This is a reflection of the different behavior observed by the different types of customers (NW customers are less patient than NE and PS customer – they probably need the service less urgently). We observed this phenomenon in Figure 15. Here, we see that the tendency to abandon maintains its order among the different types, but its magnitude is changing according to time of day, and the congestion level of the system – a fact that was not reflected in Figure 15.

Figure 30: Probability to abandon (Nov, weekdays)

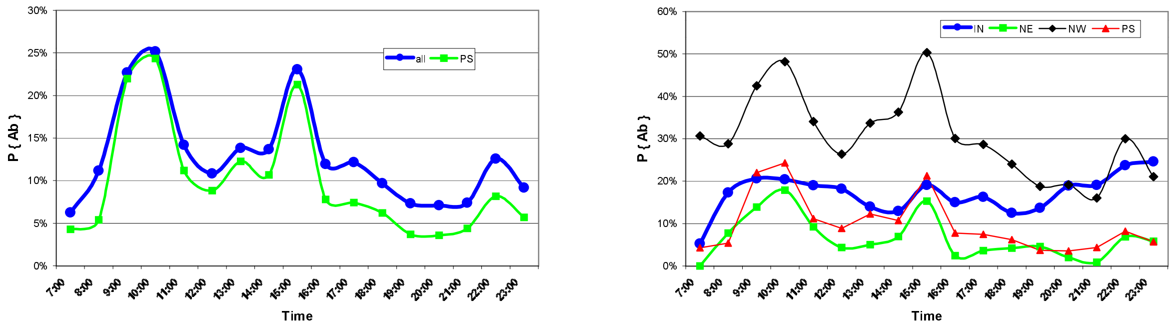


Figure 30 provides perhaps the most important display of a performance measure for a call center. Clearly, the less people abandon the higher the service level is. But this can be said about all performance measures considered so far. What distinguishes abandonment from the rest is the fact that it is the *only* customer-centric measure, through which customers inform the call center on whether the service provided is "worth its wait". (Other measures, such as delay probabilities and waiting times, are of course interesting to customers as they relate to their delay experience, but they are "objective" at the system level.)

From a behavioral point of view, the probability/fraction of abandonment should be calculated over those customers who were actually delayed at the queue (vs. over all customers, including those lucky ones who are being served immediately upon leaving the VRU.) The reason is that only the patience of those delayed is put to the test. For example, a typical figure is that 10% of the customers abandon (the other 90% reach an agent, either directly from the VRU or after waiting in queue). This 10% figure is important for the manager: it means that the system 'failed' with 10% of the callers. Now, if only 20% of the customers happen to be delayed, then actually 50% of them were impatient enough to abandon, which is more meaningful for understanding the patience characteristics of the calling population.

We now compare Figures 30 and 31. The levels of the latter are overall higher, with a mathematical explanation that is the same as for the differences between Figures 28 and 29. The peaks of both occur at the same times, but those of the latter are less pronounced. To understand this better observe that, for NW customers, the conditional probability to abandon in Figure 31 seems pretty stable over the day – about 50% of those delayed do abandon. On the other hand, the overall fraction of abandonment in Figure 30 varies between 20% to 50%; the times of peaks and valleys are close to those for the queueing time in Figure 28 – the longer the wait the more the abandonment (assuming stable patience throughout).

Further analysis of the dependence of patience on type, based on the present data-set, is carried out in [35]. The goal there is to model theoretically customers that adapt their patience to the amount of time that they anticipate to wait; their anticipation, in turn, is based on their history of visits to the call center. Empirically, such adaptivity was observed for IN customers, but not for NW.

Figure 31: Probability to abandon conditional on waiting (Nov, weekdays)

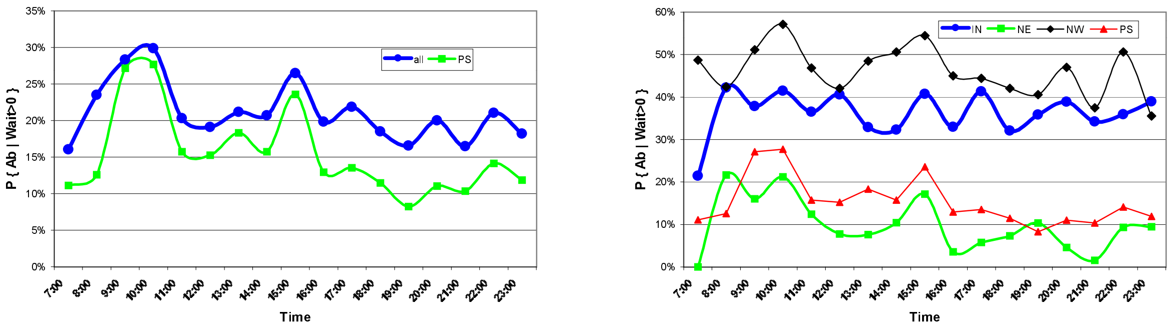
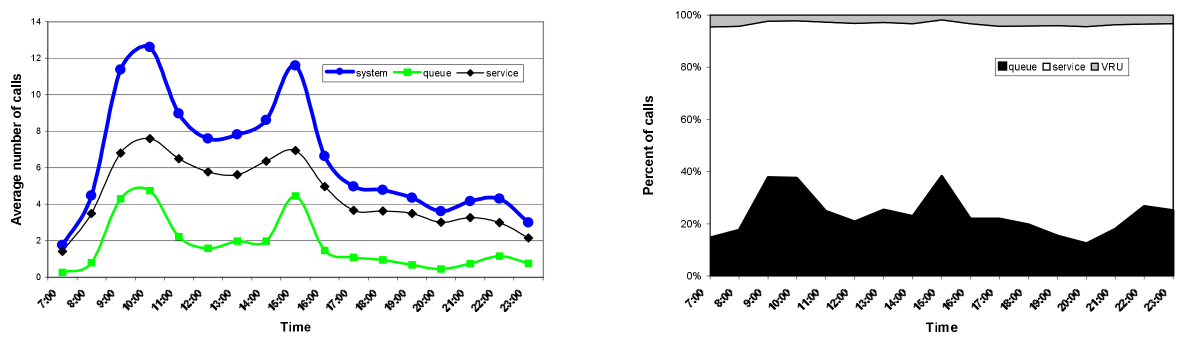


Figure 32 shows the average of the overall total number of calls present in the system, as it varies with time. A call is *in the system* at a given time interval, if it is in the VRU, queue or in service. The left plot shows the average number of calls (per hour), overall, queued or being served. (The number of calls at the VRU is negligibly smaller.) Note the correlation with the peak hours that were identified in previous figures, and that the difference between the average number of calls in the queue and in service shrinks at peaks hours. The reason seems to be the limited service capacity, determined by the number of agents working. With this number being rather constant from 8:00 to 17:00, peaks in workload translate to peaks in queue, while the number being served equals the overall number of agents (they are all busy during peak hours.)

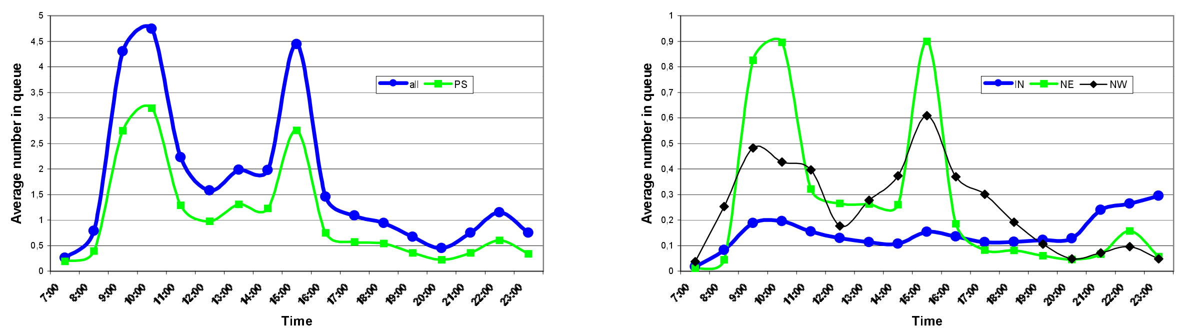
To calculate the average number of calls in the system, we first counted how many calls were in each phase (VRU, queue, service) every *minute* during the 22 weekdays. We then calculated hourly averages (over minutes) for each day, and finally averaged the daily averages over the 22 weekdays. The right-hand plot repeats the left one from a different perspective: at every given hour, we present the fraction of calls, out of the total in system, that are in each of the three phases. The black region represents queueing calls. The white region is for those in service, and the gray for VRU calls. Note that the percent of calls at the VRU is relatively small and almost constant. Note also the three peaks of the black region: as was discovered, the first two are due to "rush hour"; as will be discovered, the last one is due to reduced staffing (See Figure 35).

Figure 32: Average number of calls in system (Nov, weekdays)



In Figure 33, we examine in more detail the average number of calls in the queue: the graph for calls in the queue has a similar shape to the graph for the overall calls in the system, in the left plot of Figure 32. As calls consist mainly of type PS, this is also the situation in the queue: most calls waiting are of type PS. Note the increased difference between the graphs for 'all' and 'PS' at the times of increased number of calls of types NE and IN. Note also that the queue of IN (separate queue) reaches its maximum at the evening. Although the values in the right plot are small, they do have a large impact on workload.

Figure 33: Average number of calls in queue (Nov, weekdays)



Finally, we turn to consider the dynamics of the number of calls in the system in a particular typical day. Figure 34 shows the number of calls in the system per minute, on November 9*th*. As expected, the picture is a noisy version of Figure 32.

Figure 34: Number of calls in system (Nov 9*th*)

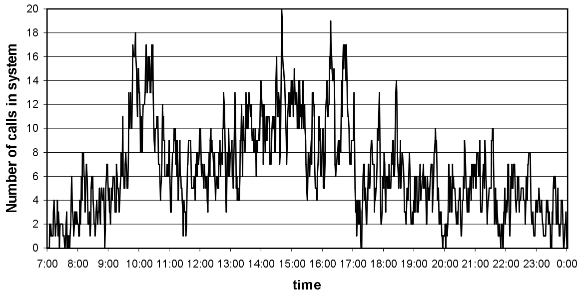


Figure 35 shows the number of calls in service, also per minute, on November 9*th*. This resembles a noisy version of the service part in Figure 32. We now use the latter figure to deduce roughly the staffing level (number of agents working), as it varies over the day. (Staffing levels constitute important characteristics that are not directly available in our data-set.) For a lower bound on the number of agents working in a given hour, let $T, T+1, T+2, \dots$ be the beginnings of consecutive minutes, and define

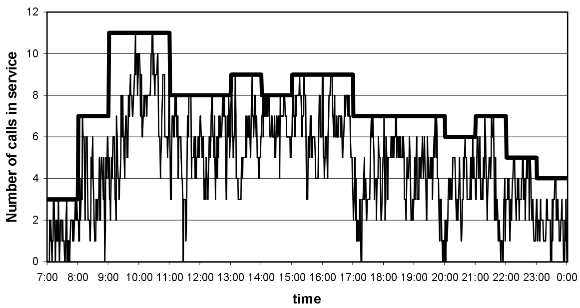
$$\begin{aligned} S_T &= \#(\text{calls which } \textit{started} \text{ service by time } T, \text{inclusive}) - \\ &\quad \#(\text{calls which } \textit{completed} \text{ service by time } T, \text{inclusive}). \end{aligned}$$

Then S_T is the exact number of calls that are served at time T (can be at the beginning, middle or end of service). Since at the time point T, an agent can provide service to only one customer, we conclude that

$$\#(\text{agents working at time } T) \geq S_T.$$

It is $\geq$, and not $=$, since not all agents assigned to the shift necessarily provide service at time T. However, it is probably true that when there are customers waiting in the queue, then $S_T = \#(\text{agents working at time } T)$. Since the number of agents working does not change by the minute, but the number of calls being served does change, we estimate the number of agents working at a given hour by the maximum of S_T over that hour (the upper envelope of S_T), which is the thick line in Figure 35. (Similar considerations could be applied over shifts.) An alternative approach to determine the number of agents working is to count the number of different agent names that are observed during an hour - we are now experimenting with this method.)

Figure 35: Number of calls in service and probable staffing level (Nov 9*th*)



9.3 Two unpredictable days

Two days in Figure 6 were much more busy than the other weekdays during 1999: on Sunday, May 23*rd*, 3064 calls arrived to the center and on Sunday, July 4*th*, 2589 calls. (The latter day has no special significance in Israel.) We now examine the effects of this unusually high number of calls on waiting and abandonment.

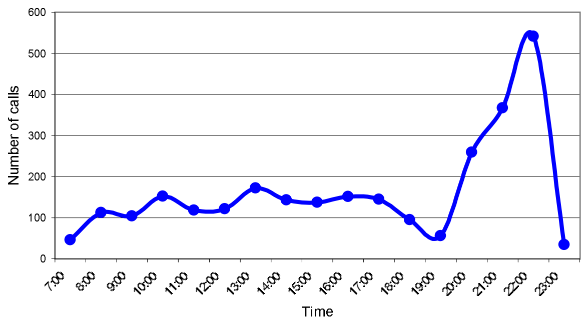
9.3.1 Sunday, May 23*rd*

Out of the 3064 calls on this day, 2752 reached the call center during its working hours, and entered queue or service with an outcome of AGENT or HANG. Of the remaining 312 calls, 258 left the system directly from the VRU, and 161 out of these 258 did so between 22:00h and midnight. We do not know if the VRU provided some clue that urged customers to abandon. (There was no service on Friday, May 21*st*, due to a holiday).

We now proceed with the 2752 calls identified above. Of these, 80.2% were type PS; 9.8% of type NW; 6.2% of type NE. There were very few calls of type IN on that day, including the evening. (Recall that the increase in IN traffic started in July.)

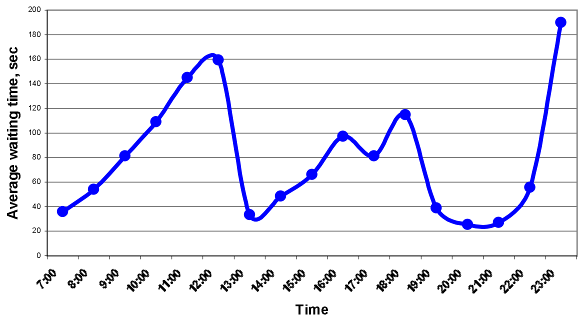
Figure 36 shows the arrival rate per hour during the day. Comparing this to the arrival rate on a typical November day (Figure 24) indicates a busy day, but not unusually so, until around 20:00h. The usual peaks during the day are obscured by the very unusual arrival pattern during the evening. About 42% (!) of all arrivals actually took place between 20:00h and 23:00h. We were not able to obtain any explanation for this.

Figure 36: Arrival rate per hour (May 23*rd*)



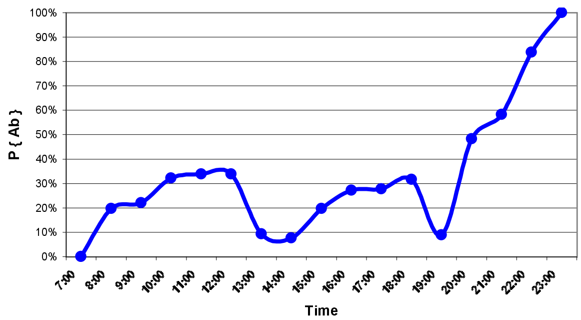
Next we examine the effect of the increased arrival rate on average queueing time per hour. Figure 37 shows it for May 23*rd*. The average queueing time is higher than the averages observed in Figure 28. The most noticeable difference is the queueing time during late evening. (Actually, between 23:00h and midnight, most of the waits were short; the average is large due to three calls that waited about 20 minutes each).

Figure 37: Average queueing time per hour (May 23*rd*)



When we compared Figures 28 and 30, we discovered that as the average waiting time increased on a typical day, so did the probability to abandon. Figure 38 shows the probability to abandon on May 23rd. Comparing it to Figure 30 demonstrates that the two mid-day peaks in Figure 38 are wider, and slightly higher than the peaks in Figure 30. However, a more interesting finding is that after 20:00h (with the increased arrival rate), the probability to abandon increased dramatically and reached 100% after 23:00h – thus, no one actually got served after this time.

Figure 38: Probability to abandon per hour (May 23*rd*)



Finally, Figures 39 and 40 show that, in the evening, not enough servers were present to cope with the high demand, and hence the long queue. Figure 39 displays the number of calls in the queue during the evening, at every minute. Note the increase in the number of calls after 23:00h, and then the decrease which is associated with the decision to abandon, made by all callers. Related to this, in Figure 40 we plot the number of calls receiving service during every minute, which exhibits a drop to 0 calls being served.

Figure 39: Number of calls in queue during evening (May $23rd$)

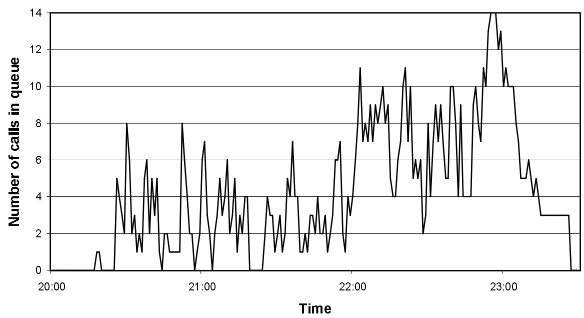
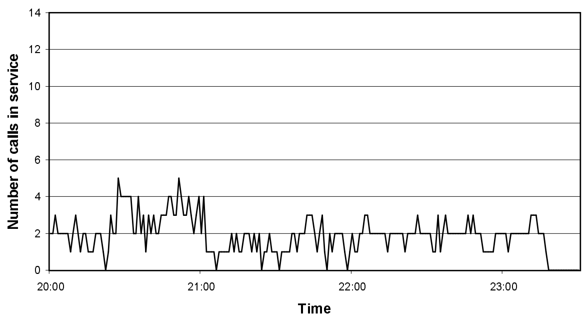


Figure 40: Number of calls in service during evening (May $23rd$)



9.3.2 Sunday, July $4th$

Out of the 2589 calls arriving to the center on this day, 2342 met the criteria mentioned for May $23rd$: 187 calls abandoned directly from the VRU, and 122 of them were after 20:20h. Most of these calls were of type NW. Actually, there was no service for type NW for nearly two hours after 20:20h, so we believe that something prevented NW callers from entering the system during this time. The distribution of types for the 2342 calls is as follows: PS – 61.2%; NW – 23.9%; NE – 7% and TT – 6.8%. Note the unusually large fraction of NW calls. This seems to be, in fact, the major difference between July $4th$ and other weekdays.

Figure 41 shows the arrival rate per hour during the day, for all calls and for types NW and NE. Note the additional peak of the arrival rate, which is due to the increased number of NW calls, and the sharp decrease in the number of calls arriving to the center after 20:00h. A possible explanation for this could be some promotion campaign, which could have increased sharply the number of NW (potential) customers.

Figure 41: Arrival rate per hour (Jul 4*th*)

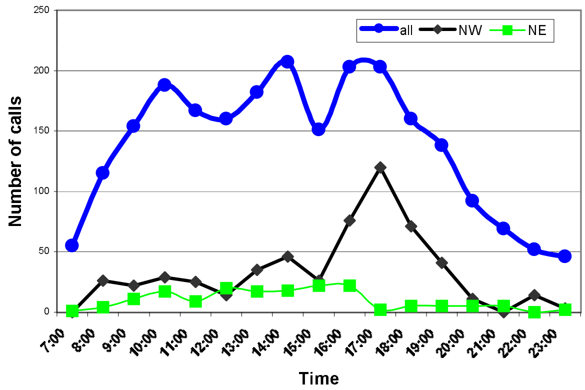
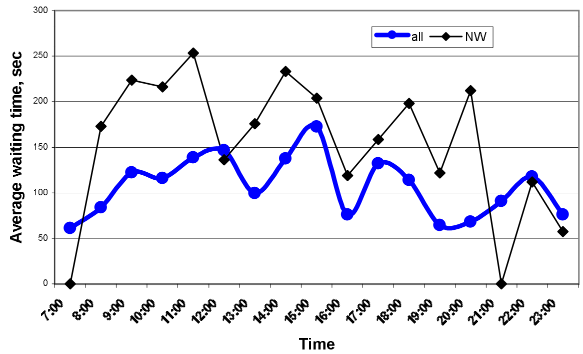


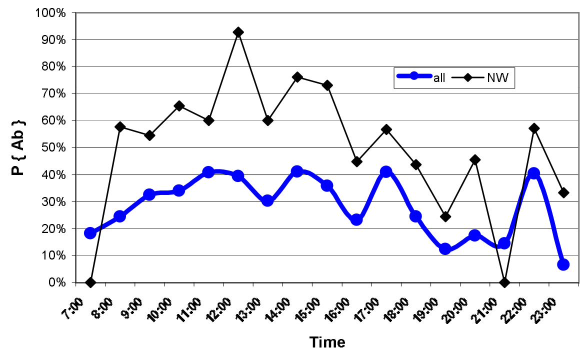
Figure 42 shows the average queueing time on this day. NW customers were exposed to longer waits (much higher than their usual queueing time; see Figure 28).

Figure 42: Average queueing time per hour (Jul 4*th*)



Did the longer queueing time of NW callers cause them to abandon more ? The answer is "yes", as demonstrated in Figure 43.

Figure 43: Probability to abandon per hour (Jul 4*th*)



Figures 44 and 45 show the number of NW calls in the queue and in service, for each minute during the day. Note the sharp increase in NW calls just before 18:00h, and the breaks in service during the day (there were very few NW calls served between noon and 14:00h, while we see that there were still calls reaching the center; note that service stopped after 20:00h). A comparison of Figures 44 and 45 indicates that, during July 4th, staffing levels were insufficient to accommodate the increased levels of NW calls.

Figure 44: Number of calls in queue during evening (Jul $4th$)

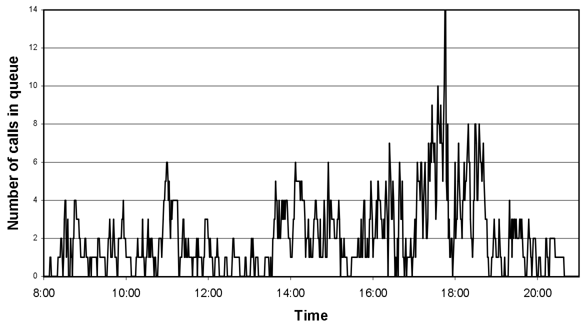
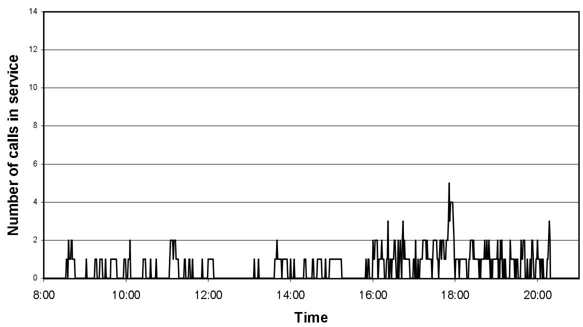


Figure 45: Number of calls in service during evening (Jul $4th$)



10 Some problematic records

In addition to unclear values of some of the fields reported in Sections 3 and 4, we note the following:

1. Calls with 0 service time, but the outcome is AGENT: Jan – 277, Feb – 85, Mar – 79, Apr – 53, May – 87, Jun – 113, Jul – 133, Aug – 125, Sep – 124, Oct – 66, Nov – 39 and Dec – 69.

2. Calls with positive service time, but outcome is HANG: Jan – 82, Feb – 97, Mar – 104, Apr – 88, May – 149, Jun – 90, Jul – 86, Aug – 126, Sep – 59, Oct – 67, Nov – 79 and Dec – 85.

11 Future research

Despite its length, our paper must be *only* the beginning. Future research that builds on the present effort could continue in many alternative directions, each important and interesting in its own right. Some such possibilities were described throughout the text. We now briefly survey a few more directions, starting with scientific research, both theoretical and empirical, and continuing with wider-scope field-studies.

11.1 Theoretical research

We barely touched the surface of the relevant *statistical* issues. Indeed, we stayed mainly at the descriptive level, leaving for the future parametric, or semi-parametric, statistical characterizations of the building-blocks of a call center (arrivals, patience, services,...). Similarly, few such characterizations have been attempted for performance measures (with notable exception being the exponential fit to the waiting times for served customers – see Figure 11). We also stopped short of any systematic analysis of inter-relations between the building blocks, and their effects on performance measures. One should now continue with applying multivariate regression, standard or censored as needed. Preliminary research already indicates that one is quickly led to unchartered territories, especially as far as censored data is concerned. Consider, merely as an example, the estimation of the hazard rate for the virtual waiting time, in Figure 14: as described there, the traditional independence assumptions of censored sampling are clearly violated, and the effects of these violations remain unclear.

Our preliminary analysis of human patience at the phone raises the need for a deeper understanding (eg. how does one rigorously describe (im)patience), as well as extensions to additional tele-services. Specifically, we propose *tele-patience* as a subject worthy of research, both empirical and theoretical, of interest to psychologists (pure and within marketing), statisticians and operations researchers. Central questions seem to be: what triggers, and what is the nature of, abandonment during VRU and Internet self-services. Here abandonment entails either a complete disengagement, or an *opt-out* where the customer seeks a human-service as an alternative to self-service. (The opposite direction is also applicable - phone customers opting out for a VRU or the Internet, for example when the alternative option requires an excessive wait for, or an expensive service by, a human operator.)

Queueing Science, along the lines described in Subsection 1.4.2, is another attractive avenue for research. This could be viewed as a natural step to follow the above-proposed statistical analysis of building blocks, and their relations with performance measures. Such relations can be used to either validate or refute the Laws of Queueing Theory, as done for example in [35].

11.2 Contact Centers

The (very near) future call center will attend to a wide customer-base. It will be connected externally to the Telephone and Internet networks and internally, through CTI, to an enterprise-wide computer database. Customers will receive multi-media information via the phone (upon request or call-backs), a Web site, e.mail or fax. Future ACD's will increasingly route requests to electronic agents — yet, we believe, *the human-service* is with us to stay. Redoing the present study at a multi-media (perhaps also multi-site) call center, or contact centers as they have come to be known, is an important natural next step.

11.3 Data integration

As indicated in Subsection 1.2, the integration of ACD, CTI and Psychological data, at the individual call level, is of utmost importance if one wishes to correlate business success with service-quality. For example, testing whether short calls are those that generate most revenues, or perhaps the longer ones do. Or more generally, the goal is to identify the performance characteristics of successful business transactions, in particular best-customers' and best-agents' characteristics.

11.4 The Ideal

We envision a data-repository that is continuously fed by many call centers, of varying types (business, information, emergency,...) that are telephone- and/or Internet-based. (To our understanding, Lucent's Definity, and very likely products of other companies as well, have the capabilities to create such repositories). The collected data is to be continuously and automatically analyzed, from both operations and marketing perspectives. Then the data is to be both archived and fed back to the originating call centers, who would use it (through visualization tools) to support ongoing operations, as well as tactical and strategic goals. Little imagination is required for appreciating the value of such a data-base. As a start, its developer could become a benchmark that sets industry standards, as far as customer-service quality and call-center efficiency are concerned. As already mentioned, such a data-base would enable the identification of success-drivers of call-center business transaction.

This ideal can be achieved through the following steps:

(A) Expand the present step: Rigorously analyze operations data, in order to learn and document operations of call centers.

(B) Rigorously analyze and record marketing data. This direction seems to be related to "data mining" research, that is gaining popularity. (Data mining research is presently carried out by two communities: statisticians and computer scientists. Our understanding is that they do it rather separately, but we are convinced that success would require a truly joint effort).

(C) Combine (A) and (B), namely combine operations and marketing data. The goal here is to develop a model (descriptive, prescriptive; empirical, analytical;,...) that relates QOS (= Quality of Service) with the Bottom Line (Profit from Service).

For example, a long service is considered inefficient, but if such services tend to lead to large profits, they should not be discouraged. Similarly with respect to repeat callers (retrials.)

For example, long waits can be accompanied by marketing information which, in turn, could actually have a positive contribution to profit (but maybe not).

For example, information about expected waiting time, or the option of leaving a recorded message to the call center, is helpful in terms of QOS but possible harmful in terms of profit.

etc.

(D) Continue (C) with Psychological data: sources here are typically customers, agents and managers surveys, which greatly vary but are prevalent in decently-run call centers. (One untapped source for survey data is the VRU, which is unique for telephone operations in the following sense: after each call, the agent can ask the customer to answer a couple of questions, regarding QOS. This survey can actually be carried out through a VRU, with the option of direct supervisor's support in cases of "hard-feelings"),

(E) An important trend in call centers is the incorporation of Internet-based services. It is easy to create analogies of all the above, with respect to Internet data, in case of a pure-Internet call center. The issues become even more interesting when the Internet is integrated with a telephone-based call center. Then we have two sources of Operations data - ACD records and Internet log-files.

We believe that it might be actually possible to delve directly into (E), in parallel to our efforts with (A), namely identify an Internet-based call center that allows access to its log-file. Our assessment is that one is more likely then to have access to individual=call level data here.

References

[1] R.N. Anthony. *Planning and Control System: A Framework for Analysis.* Harvard University Press, 1965.

[2] J. Anton. The Purdue benchmark research. `http://www.e-interactions.com`, WEB site.

[3] D. Bertsekas and R. Gallager. *Data Networks.* Prentice Hall, 2nd edition, 1992.

[4] S. Borst, A. Mandelbaum, and M. Reiman. Dimensioning of large call centers. Preprint, 2000.

[5] A.J. Brigandi, D.R. Dargon, M.J. Sheehan, and T. Spencer III. AT&T's call processing simulator (CAPS) operational design for inbound call centers. *Interfaces*, 24:6–28, 1994.

[6] E.S. Buffa, M.J. Cosgrove, and B.J. Luce. An integrated work shift scheduling system. *Decision Sciences*, 7:620–630, 1976.

[7] J.A. Buzacott and J.G. Shanthikumar. *Stochastic models of manufacturing systems.* Prentice Hall, 1993.

[8] D. Cox. Regression models and life tables. *Journal of the Royal Statistical Society, B*, 34:187–220, 1972.

[9] A.K. Erlang. The theory of probability and telephone conversations. *Nyt Tidsskrift Mat. B*, 20:33–39, 1911.

[10] A.K. Erlang. Solutions of some problems in the theory of probabilities of significance in automatic telephone exchanges. *Electroteknikeren (Danish)*, 13:5–13, 1917. English translation: 1918 P.O. Elec. Eng. J. 10, 189–197.

[11] O. Garnet and A. Mandelbaum. *An Introduction to Skills-Based Routing and its Operational Complexities.* Teaching note sponsored by the Fraunhofer IAO Institute, Stuttgart, Germany, June 1999.

[12] O. Garnet, A. Mandelbaum, and M. Reiman. Designing a telephone call-center with impatient customers. Submitted for publication, 1999.

[13] L. Green and P. Kolesar. Testing the validity of a queueing model of police patrol. *Management Science*, 3:127–148, 1989.

[14] D. Gross and C.M. Harris. *Fundamentals of Queueing Theory.* Wiley, 3rd edition, 1998.

[15] R.H. Hall. *Queueing methods for services and manufacturing.* Prentice Hall, 1991.

[16] W. Härdle. *Applied nonparametric regression.* Cambridge University Press, 1990.

[17] C.M. Harris, K.L. Hoffman, and P.B. Saunders. Modeling the IRS telephone taxpayer information system. *Operations Research*, 35:504–523, 1987.

[18] T. Hastie and R. Tibshirani. *Generalized additive models.* Chapman and Hall, 1990.

[19] R. Herman. Technology, human interaction, and complexity: reflections on vehicular traffic science. *Operations Research*, 40:199–212, 1992.

[20] W.J. Hopp and L.M. Spearman. *Factory Physics.* IRWIN, 1996.

[21] Help-Desk Institute. Annual reports by the Help Desk Institute (HDI), survey.

[22] J. Kalbfleisch and R. Prentice. *The statistical analysis of failure time data.* Wiley, 1980.

[23] L. Kleinrock. *Queueing Systems*, volume 1 and 2. Wiley, 1975.

[24] C. Kooperberg, C. Stone, and Y. Truong. Hazard regression. *Journal of the American Statistical Association*, 90:78–94, 1995.

[25] A.M. Lee. *Applied queueing theory.* MacMillan, 1966.

[26] Call Center Magazine. `http://www.callcentermagazine.com`, WEB site.

[27] R. Miller. *Survival analysis.* Wiley, 1981.

[28] C. Palm. Methods of judging the annoyance caused by congestion. *Tele*, 4:189–208, 1953.

[29] J.W. Roberts. Recent observations of subscriber behavior. In *9th International Tele-traffic Conference (ITC-9)*, Torremolinos, 1979.

[30] B. Silverman. *Density estimation.* Chapman and Hall, 1986.

[31] Call Center Statistics. `http://www.callcenternews.com/resources/statistics.shtml`, WEB site.

[32] D.Y. Sze. A queueing model for telephone operator staffing. *Operations Research*, 32:229–249, 1984.

[33] W. Venables and B. Ripley. *Modern applied statistics with S-plus.* Springer, 3rd edition, 1999.

[34] R.W. Wolff. *Stochastic Modeling and the Theory of Queues.* Prentice Hall, 1989.

[35] E. Zohar, A. Mandelbaum, and N. Shimkin. Adaptive behavior of impatient customers in tele-queues: Theory and empirical support. Preprint, 2000.