The relationship between the waiting crowd and the average service time

Oliver Handel and Andre Borrmann

Abstract In this paper, the relationship between the waiting crowd and the service time – the average duration to serve one single customer – in the context of vendor stands (e.g. food stands, concession stands or kiosks) is evaluated. Drawing from traditional analytic queuing theory, a distribution function for the service time that remains steady is generally used. This steady state assumption is questioned in this paper by using computer simulation, empirical observation and qualitative reasoning. On the one hand, the impact of the amount of people waiting on the average duration of service time is examined. On the other hand, the effects of crowding on the choice of a customer are evaluated as well. Within this context different causal feedback relationships are identified that are expected to be of fundamental importance. The paper concludes that for the endogenization of the service time, the incorporation of these feedback relationships is key to obtain more accurate results.

1 Introduction

While the time a person spends waiting in a queuing situation is commonly overestimated, the satisfaction tends to decrease the longer the waiting time is perceived [1, 2]. To avoid dissatisfaction of the people getting served, two different approaches can be distinguished. The first approach aims to manage the actual waiting time by predicting the demand and to provide a sufficient amount of servers to customers. But as services cannot be inventoried [3] and the demand for the service is hard to predict, waiting is often unavoidable if the cost of servers are not neglected. Various

Oliver Handel
Technische Universitt Mnchen, Arcisstr. 21, 80333 Munich, Germany.
e-mail: oliver.handel@tum.de

Andre Borrmann
Technische Universitt Mnchen, Arcisstr. 21, 80333 Munich, Germany.
e-mail: andre.borrmann@tum.de

approaches from operational management research aim to minimize the actual waiting times, but because of the difficulties in providing the right amount of servers in every situation and to control therefore the actual wait duration, another approach aims to reduce not the actual waiting time, but the perceived waiting time by focusing on different characteristics of the service environment that affect time perception and thus make the waiting experience for the people waiting less dissatisfying [4]. Hence, this approach tries to influence the subjective perceived waiting time by drawing from theories from sociology, psychology and marketing. The perception of wait time and service satisfaction were discussed by [4, 5, 6] among others. Essential in this discussions are the degree of social interaction and distraction from the situation in the filled time gap [5], considerations about social justice (first-in-first-out-principle) and elements from the service environment (lightning, temperature, music, color and furnishings) [4] that influence the perceived waiting time.

Although it is smart aiming to make the duration of the wait as comfortable as possible and thus to decrease the perceived waiting time and to avoid that the use of a service is overshadowed by the frustration of a perceived long wait, the even smarter way is to better optimize the actual waiting time and to overcome some obstacles in doing so. One key issue is the service time, defined as the duration to serve one single customer. Commonly the service time is seen as an input value that needs to be empirically collected, statistically aggregated and then inserted in the evaluation method as a fixed parameter or distribution function that remains the same over the whole time span. In this paper, the static service time assumption is rejected and the dynamic nature of the service time is pushed to the fore. Collected empirical data in the context of vendor stands at a music festival provides evidence that an endogenization of the service time variable is necessary to increase the forecast accuracy for the length of waiting queues within the simulation. Before results from simulation are presented, analytic queuing theory is discussed in the next section with the outlook that there is need for simulation in this context.

2 Analytic Queuing Theory and Kendall's notation

Analytic queuing theory aims to mathematically describe performance functions (e.g. average waiting time of a customer or server utilization rate) of different queuing systems [7]. In this domain, queuing is not limited to queuing pedestrians, but also includes other queuing situations. Kendall's notation [8] prevailed to classify queuing systems. In this notation a queuing system is defined by a row of different letters: $A/S/c/K/N/D$ (A = arrival process / S = service time distribution / c = number of servers, K = capacity of the system / N = population size / D = service discipline). In case of pedestrian queuing, the short form of the Kendall notation $A/S/c$ can be generally used, because K and N are commonly assumed to be infinite and a FIFO service discipline is expected. A and S describe then different distribution functions, such as the Poisson, Degenerate, Erlang or Phase-type distribution. An abbreviation for each distribution function is used, plus the number for

the amount of servers to define the queuing system. After having defined the queuing system, different performance functions can be specified. From a customer-focused perspective, the number of waiting customers and the average waiting time can be examined. From a server-focused perspective, performance functions such as the idle and busy time of a server or the utilization of the server can be measured.

Over the past hundred years, several scholars have contributed to solve different queuing systems analytically. An overview about which queuing systems are analytically solved can be found in [9]. It became a competition in probability and queuing theory to solve these commonly called *waiting time problems* [10], because – from a practical perspective – the only possibility to describe the performance of different pedestrian queuing systems was in times without computational simulation the analytic approach beside empirical observation. In other words, scholars who only had the analytic approach as their hammer saw every *waiting time problem* as a nail, nevertheless which limitations this approach embrace. Although it is of scientific value to solve queuing systems mathematically, some managerial implications remain especially in the domain of pedestrian queuing situations. Firstly, Kendall's notation includes the steady state assumption, as the arrival process and the service time distribution remain constant. Therefore, the formulas are helpful to get some quick benchmarks on how the queuing system would perform under the given narrow model boundary constrains, but taking into account more realistic scenarios with variations of the arrival pattern, the analytic approach is of limited help, if an overall evaluation is necessary and thus there is need for simulation. Simulation enables to take into account more dynamic arrival patterns or variations of the service time or to endogenize these key factors. Finally, in respect of the research topic, Kendall's notation may have contributed to the erroneous assumption that the amount of waiting people and the service time distribution are in each case two independent variables.

3 Simulation of pedestrian queuing

Simulation has several advantages. First of all, different from the analytic approach, simulation enables the generation of benchmarks for more complex queuing situations (e.g. oscillating batch arrivals). Secondly, in case of pedestrian queuing situations, the physical layout of the queuing environment – the servicescape [11] – can be taken into account, leading to minor delays, if a walking distance from the end of the queue to the server is necessary. And thirdly, simulation enables to endogenize key factors and therefore to push the model boundary forward.

The Java-based software Anylogic [12] is used here for the simulation of pedestrian queuing. In Figure 1, a snapshot of a *M/M/5*-queuing-system-simulation is shown. The set-up consists of five servers and a single queue in front of the servers. The arrival process and the service rate are Poisson distributed. On the right hand side of the figure, the amount of waiting customers are depicted in the upper diagram and the utilization rate of the servers are shown over time. In accordance with

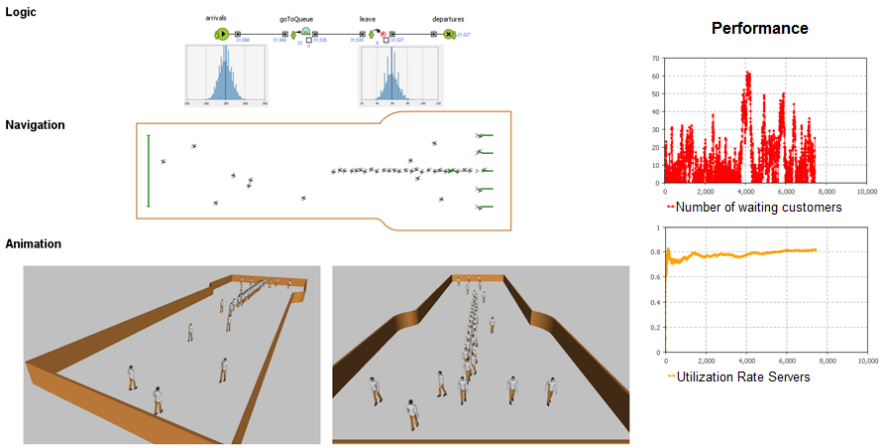

Fig. 1 Pedestrian queuing simulation with Anylogic

the amount of five servers and the given service time distribution with a mean value of 41,6 seconds and an arrival rate with a mean value of 250 arrivals per hour, a utilization rate of around 80 % in steady state can be measured under the boundary condition that the amount of waiting people has not an effect on the service time.

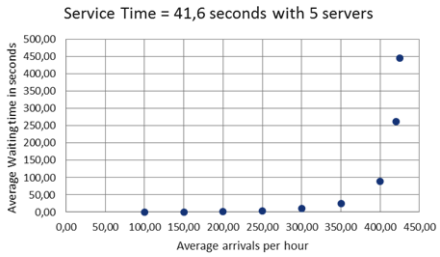

Fig. 2 The relationship between the arrival rate and the average waiting time. The blue dots are the measured average waiting times as results from the simulation.

If the arrival rate is increased step-wise from 100 to 450, Figure 2 indicates that there is some qualitative moment of change where the waiting time increases drastically and leads to infinite waiting times. The diagram shows that if the number of arrivals increases from 350 arrivals per hour to 400 arrivals per hour (15 % increment), the waiting time increases from 24 seconds to 90 seconds (375 % increment) and goes from there on in the steady state quickly to infinity, because the waiting lines get endless long. The point where the queuing system cannot cope with the number of arrivals and the queues get endless long can be called a tipping point. The occurrence of tipping points in queuing systems has managerial implications, as the aim is to keep the queuing system away from the tipping point. The good news is that there are generally feedback effects leading to an increase of the maximum

throughput of the queuing system, if waiting times get long. This form of systemic self-organization will be discussed in the next section and results into the dynamic service time assumption.

4 Empirical assessment and findings

For the set-up of a microworld of an urban event case study, empirical data collection was conducted to gather essential input values such as the average time it takes to serve one customer. The data acquisition was carried out at a music festival in 2014 and 2015 in Garching, Germany. Video cameras have been used to assess the waiting crowd in front of several vendor stands (outdoor bars and food stands), mobile toilets and the entrance facility, and to get empirical values for the service times respectively durations of use. Post-evaluation of the primary video data was conducted to collect the secondary data material. A detailed report about the 2014 data collection and the spots observed can be found in [13].

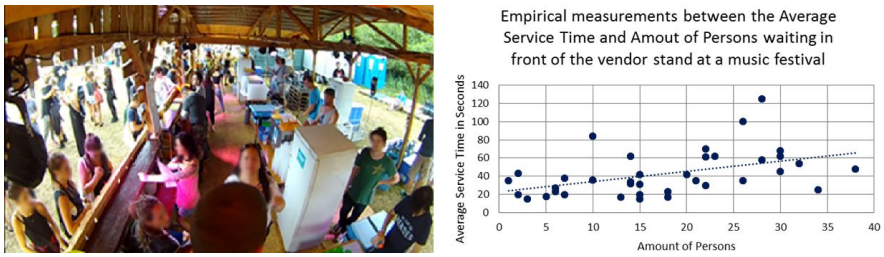

Fig. 3 Empirical collection of service times and the amount of people waiting at the same moment in front of a vendor stand at a music festival in Garching, Germany.

In Figure 3, the left picture shows exemplarily such an observed bar with different operational staff from the inside. On the right side of the figure, different measured data values for the time length of the collected service times are marked together with the information how many people waited in front of the vendor stand at this moment in time. Surprisingly there exists a positive correlation between both features (indicated by the linear trend line), i.e. the more persons waiting, the longer the average service time. A causal explanation about this counterintuitive finding is shown with the next figure.

Figure 4 shows a dynamic hypothesis in form of a causal loop diagram. Through qualitative reasoning three main mechanisms (efficiency increase, grouping and postponement effect) are identified to be important in this context and are summarized in the figure. The first assumption is that an increase of waiting people in front of the vendor stand leads to an increased pressure on the employees and causes an efficiency increase that reduces the service time per customer and thus the amount of people queuing. This effect is limited by an efficiency maximum. But beside this

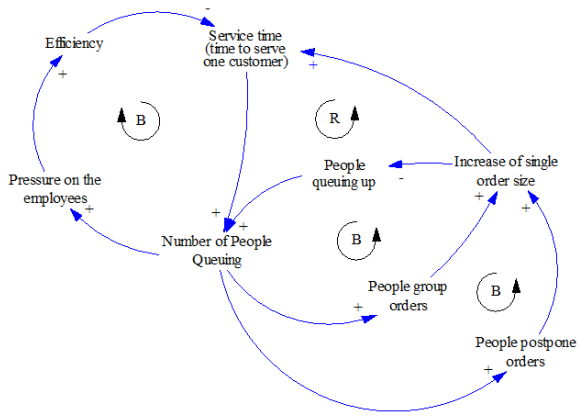

Fig. 4 The relationship between the number of people queuing and the service time in form of a causal loop diagram.

effect, if more and more people queue up, two other stronger mechanisms also begin to operate – so the second and third assumption – leading to the increase of the service time. One the one hand, if many people queue up, people start to group orders by asking a friend to bring along something for them and on the other hand, people start to postpone their orders, leading both to an increased single order size. If a fixed total order volume is assumed, the bigger the single orders are, the longer it takes to serve one customer (increase of the service time), but also the less amount of people need to be served. This effect is twofold counterintuitive, because firstly the service time increases, but secondly – at the same time – the total order volume throughput increases. In compliance with the assumptions made, this finding is an example of positive self-regulation within a system.

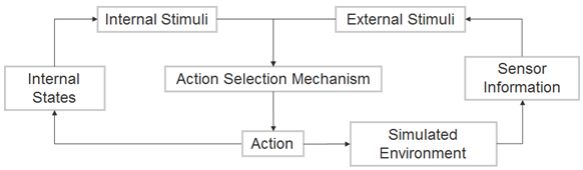

Fig. 5 A generic concept of an action selection mechanism. Internal and external states affect decision-making.

To embed these mechanisms in a simulation, it is necessary to feedback from the amount of waiting people on the service time directly and on agent level on the decision architecture that defines under which conditions people decide to queue up, group and postpone orders. To make the simulation even more accurate, it is furthermore necessary to incorporate the grouping of orders effect as well. In Figure 5, the generic concept of a decision architecture is shown. While the amount of waiting

people is in this form a sensory information coming from the simulated environment and affects the dynamic decision-making of the agents and thus generate the grouping and the postponement of orders, the dynamic change of the service time will affect the simulated environment directly. More details about how crowding affects the choice of a customers to queue up are summarized in the next section.

5 Crowding, customer choice and queue shape selection

In accordance with the context, different shape formations of the waiting crowd are possible. On the one hand, it is possible that very well organized queues are formed in the shape of a single or multiple queue structure. On the other hand, without according barriers it is often the case that an unorganized densely packed waiting crush forms. In Figure 6, the implementation of different queuing models are depicted based on the Anylogic software. These models allow to embed the different possible queuing formations in the simulation. How these different formations influences the choice of further customers to queue up are described in the following.

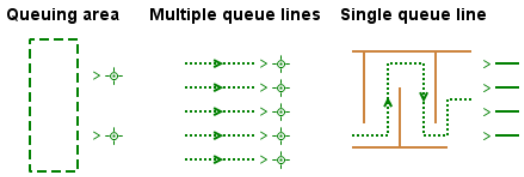

Fig. 6 Three different queuing models: 1. Service points with a queuing area. 2. Service points with separate queue lines. 3. Several service lines with one single queue line.

Density and crowding are related to each other, but a differentiation is necessary [14]. Density refers to the physical condition – according to the spatial parameters [15] and crowding is more related to the unpleasant feeling of an individual in terms of the control perception to move freely in the environment. [16] used the term perceived control as an intervening variable between density and crowding that influences the behavior of an individual. Drawing from the work of [14], perceived crowding affects negatively the emotional and behavioral responses of an individual and thus may hinder individuals to queue up, so the assumption. [17] argued that high-density conditions affect the risk-potential, as the velocity decreases in dense crowds. Therefore the question which queue shape to foster from a managerial perspective are quite straightforward to answer. The more perceived control the individual will have, the less likely that negative emotional responses are expected by the individual and the more likely that the individual not avoids to queue up. Therefore, the first choice should be to avoid the formation of an unorganized waiting crowd, and secondly in accordance with [18] a single queue structure is more preferred than a multiple queue structure, because of fairness and predictability considerations.

6 Conclusion

In this paper, the relationship between the waiting crowd and the average service time have been evaluated. In the beginning, the classical analytic approach to solve queuing systems was discussed. Drawbacks of this approach have been mentioned, such as the non-endogenization of essential key parameters and based on these considerations the assumption was made that the analytic approach strengthened the error-prone perception of a steady service time. From this point of view, the need of simulation was emphasized as simulation enables to take into account the physical layout of the queuing situation and to incorporate essential feedback effects to endogenize variables of the queuing system and therefore to see the queuing system not as isolated situation. Simulation of pedestrian queuing demonstrated the occurrence of tipping points, as the thresholds where the queuing system can either cope with the amount of arrivals or the waiting queues get endless long. Based on these considerations, several feedback effects between the waiting crowd and the average service time that occur in the real-world and prevents the system from going beyond the tipping point, have been discussed. The efficiency increase, the postponement and the grouping effect was introduced based on a causal loop diagram and the incorporation of these effects into simulation was discussed. Finally, the effect of crowding on customer choice was elaborated and different queuing models have been introduced.

References

1. K.L. Katz, B.M. Larson, R.C. Larson, Sloan Management Review **32**(2), 44 (1991)
2. M. Groth, S.W. Gilliland, Journal of Quality Management **6**(1), 77 (2001). DOI 10.1016/S1084-8568(01)00030-X
3. V. Zeithaml, A. Parasuraman, L. Berry, The Journal of Marketing (1985)
4. J. Baker, M. Cameron, Journal of the Academy of Marketing Science (1996)
5. S. Taylor, Journal of Marketing **58**(2), 56 (1994). DOI 10.2307/1252269
6. D. Maister, The Service Encounter (2005)
7. D. Gross, C.M. Harris, *Fundamentals of queueing theory* (John Wiley & Sons, 1988)
8. D. Kendall, The Annals of Mathematical Statistics (1953)
9. L. Kleinrock, *Queuing systems* (Wiley, 1975)
10. L. Takács, Acta Mathematica Academiae Scientiarum Hungaricae **6**(1-2), 101 (1955). DOI 10.1007/BF02021270
11. M. Bitner, The Journal of Marketing (1992)
12. A. Borshchev, p. 612 (2013)
13. D.H. Biedermann, F. Dietrich, O. Handel, P.M. Kielar, M. Seitz, Using Raspberry Pi for scientific video observation of pedestrians during a music festival. Tech. rep., Technische Universität München, München (2015)
14. M. Hui, J. Bateson, Journal of Consumer Research (1991)
15. D. Stokols, Psychological review **79**(3), 275 (1972). DOI 10.1037/h0032706
16. D. Schmidt, J. Keating, Psychological Bulletin (1979)
17. O. Handel, D.H. Biedermann, P.M. Kielar, A. Borrmann, Transportation Research Procedia **2**, 669 (2014)
18. A. Rafaeli, G. Barron, K. Haber, Journal of Service Research **5**(2), 125 (2002). DOI 10.1177/109467002237492