International Journal Of Engineering Research And Development
e- ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com
Volume 20, Issue 9 (September, 2024), PP 185-192

Emergency Response System Call Center Staffing Characterization with Integrated Demand using Stochastic Simulation

Javier Holguín-De La Cruz[1]
[1]*Department of Industrial and Manufacturing Engineering, Institute of Engineering and Technology*
Universidad Autónoma de Ciudad Juárez,Ciudad Juárez, Chihuahua, MEXICO
Corresponding Author: Javier Holguín-De La Cruz

ABSTRACT *National insecurity perception levels in Mexico present an average of 74.6 % among citizens 18 years old and older according to INEGI ENVIPE national poll (INEGI, 2023). This average is obtained from answers provided by men with a value of 70%, and women with a value of 78.6%. This national poll reports the lowest average of 72.3 % in 2013, and the highest value of 79.4% in 2018. From these statistics it is observed that at least from 2013 to the present, an average of 7.53 out of 10 inhabitants perceive the public environment as insecure. These abnormally and tremendously high statistics do not assist in reaching acceptable and peaceful living standards. Public safety Emergency Response Systems are the number one resource that society relies on to maintain a peaceful environment and the state of law. In these systems, minimizing response time is crucial to deter and control crime. As part of these emergency systems, their call centers, which process the calls for service, demand an ideal allocation of personnel or agents responsible to expedite calls and dispatch units. As an effort to optimize or identify an ideal allocation of call answering agents, we characterize ringdown time as a function of the allocated number of call answering agents for an integrated demand for service including non safety related calls and expected false calls using stochastic discrete simulation.*

KEY WORDS *Emergency Response System, Response Time, Ringdown Time, Call Center, Staffing Allocation*

Date of Submission: 05-09-2024 Date of Acceptance: 18-09-2024

I. INTRODUCTION

A further analysis of the higher percentage of insecurity perception of women compared to the percentage referred by men according to the Perception and Victimization National Poll (ENVIPE) 2023 (INEGI, 2023), reflects that since 2013, a considerable difference of an average of 6.57% higher percentage is observed. These statistics also reflect that on average 7.17 and 7.83 out of 10 men and women respectively, perceive an insecure social environment. Moreover, it is clearly established that women perceive a higher insecurity level than men, and that this could be an indicative that they possibly need improved strategies to better protect them in our society.

Emergency Response Systems (ERS) integrate several organizations providing emergency services including police, fire, and medical (Jennex, 2007). According to Piyadasun et al. (2017) and van Barneveld et al. (2018), the key performance parameter of all ERS is *response time*, which is measured from the time a call is answered to the time a unit arrives to the location where assistance is demanded (D'Amico et al., 2002). Nevertheless, there is another component of the *response time* usually not considered as part of the waiting time, identified as *ringdown time*, which is defined by the U.S.A. Department of Justice, Bureau of Justice Statistics (Yung and Dayharsh, 1980) as the time a phone rings before it is answered.

Both of these times are required to be considered when a Public Safety Answering Point (PSAT), which is part of the 911 Emergency Response System in the United States of America, is being designed to serve a given geographical area with an expected call for service demand (Yung and Dayharsh, 1980). Based on the call for service demand of a given PSAT, and the operating policies for the quality of service considering ideal *ringdown*, *response times* and the level of service to be offered defining the percentage of calls to be processed under this performance parameters, the call center of the PSAT configures the ideal number of answering agents and associated infrastructure (Yung and Dayharsh, 1980). As a reference of the ideal *ringdown time*, in the United States of America, the Department of Justice, Bureau of Justice Statistics, states that 90% of the calls must be answered within 10 seconds (Yung and Dayharsh, 1980). In the literature, this problem is recognized as a staffing problem to configure the ideal allocation of call answering agents in call centers subject to restrictions including *response time* and *ringdown time* (Li et al., 2019; Champanit and Udomsakdigool, 2020).

In our research we address the problem related to the configuration of the ideal number of answering agents through characterization in the call center of the PSAT in a 911 Emergency Response System in México.

We considered total demand for service corresponding to one out of nine police districts in a large city, coming from all corporations including police, fire and medical emergency services. In addition, this research also includes an estimated percentage based on recent publications, of calls of demand for service, which were false, not completed, hung up or abandoned. A previous research evaluated only safety related calls for service.

Our research generates stochastic simulation models of the call center processes of answering and dispatching and evaluates the queueing processes of incoming calls as an equivalent of the *ringdown time* incoming calls experience before being answered. Data was obtain from historical operation of 552 consecutive hours and results were evaluated and contextualized based on the population increase to the present time as a predictive factor of demand behavior of calls for service.

II. LITERATURE REVIEW

The concept of the *ringdown time* or *call answering interval* associated to an Emergency Response System is referred in the United States of America by the National Emergency Number Association (NENA, 2020) and the Department of Justice, Bureau of Justice Statistics (Yung and Dayharsh, 1980) as the time interval between the call rings at the PSAT and when the call is answered. A diagram that presents the call processing at the Call Center at the PSAT in the Emergency Response System is presented in Figure 1 (aNENA, 2020).

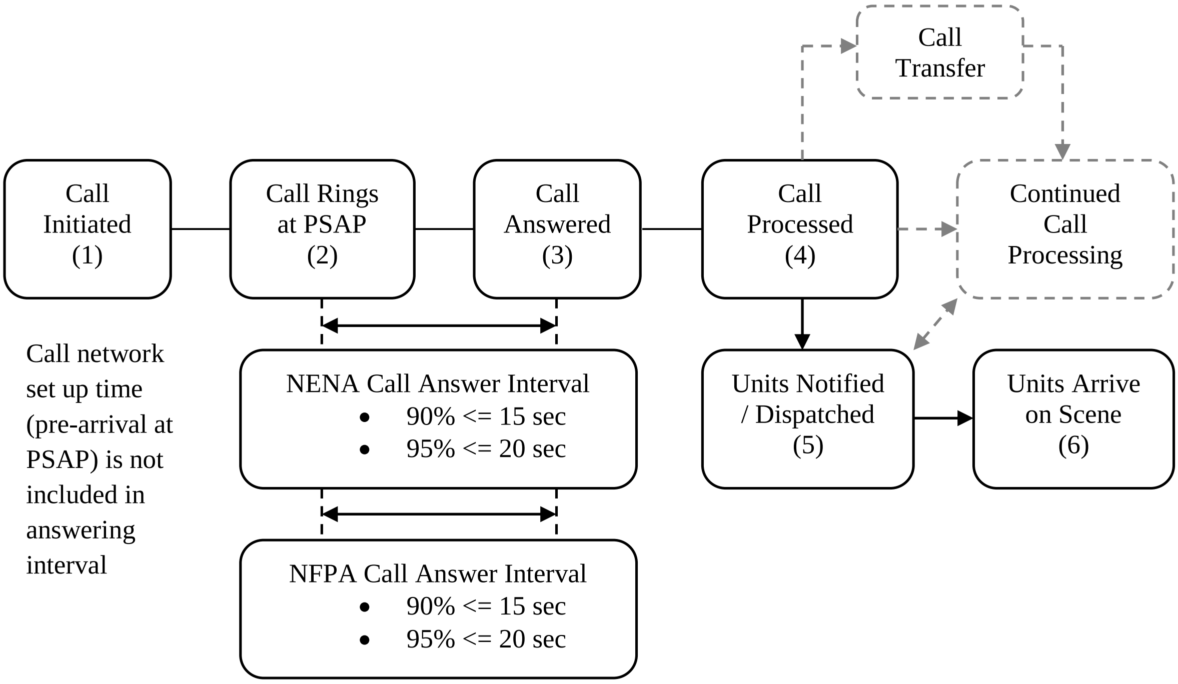
Call Transfer
Call Initiated (1)
Call Rings at PSAP (2)
Call Answered (3)
Call Processed (4)
Continued Call Processing
Call network set up time (pre-arrival at PSAP) is not included in answering interval
NENA Call Answer Interval
• 90% <= 15 sec
• 95% <= 20 sec
Units Notified / Dispatched (5)
Units Arrive on Scene (6)
NFPA Call Answer Interval
• 90% <= 15 sec
• 95% <= 20 sec

Figure 1: NENA call answering interval for 9-1-1 calls (NENA-STA-020.1-2020, 2020)

In Figure 1, it can be observed that the *call answer interval* or *ringdown time* as referred by Yung and Dayharsh (1980), is obtained by subtracting the continuous time reading in Step 2 identified as “Call Rings at PSAP”, from the time reading in Step 3 identified as “Call Answered.” This figure also illustrates the required performance parameters for both, the National Emergency Number Asociation (NENA) and the National Fire Protection Association (NFPA), which establishes that 90% of the calls must be answered in 15 seconds or less and 95% of the calls must be answered in 20 seconds or less.

The application of stochastic simulation to evaluate and improve Emergency Response Systems is widely utilized. Examples of these applications are provided by Ye et al. (2021), Huang (2015), Hatami-Marbini et al. (2022). A literature review conducted for the analysis and optimization of call centers in Emergency Response Systems or in the service industry integrates publications by Brooks et al. (2011), Conley and Grabau (2013), Ibrahim et al. (2012), van Buuren et al. (2015; 2017), L’Ecuyer et al. (2018), Li et al. (2019), Seada and Eltawil (2015), Steinmann and de Freitas Filho (2013), Ta, T.A. (2021), Liao et al., (2009), Yu et al., (2018).

van Buuren et al. (2017) present a simulation model of a call center for a medical emergency service where answering agents may have different functions. They created three discrete event simulation models to represent: (1) function differentiation in two agent classes, (2) only one common class for all agents, and (3) a combination of the two models described in (1) and (2). The authors established that the objective of their research is to provide insight in addressing strategic issues such capacity and work force planning, which are required in all other emergency call centers.

In the United States of America, the National Emergency Number Association (NENA) has identified three different categories for uncompleted calls to the 911 Emergency Response System (bNENA, 2020). The descriptions of these categories are:

> "**silent** 9-1-1" call is defined as *"someone has dialed 9-1-1, the call has successfully passed through the 9-1-1 network and has been answered by a 9-1-1operator. Aside from the 9-1-1 operator's, no voice communication is heard on the initiating caller's end of the emergency call."*
>
> "9-1-1 **hang-up**" call is defined as *"someone, either through malicious intent or accidental occurrence, has dialed 9-1-1, the call has passed through the emergency network and has been answered by a 9-1-1 operator. The initiating caller has hung up prior to the 9-1-1 operator answering the call."* This type of call is very similar to an abandoned 9-1-1 call.
>
> "**abandoned** 9-1-1" call is defined as *"someone has dialed 9-1-1 and all available operators are busy. The call is placed into queue for answer. Rather than wait for an available operator, the caller elects to hang-up prior to the 9-1-1 call being answered by an available 9-1-1 operator."* Generally, an "abandoned" call is more often associated with call centers that have an automatic call distribution (ACD) call delivery scheme.

According to the Executive Secretariat of the Secretary of Security and Citizen Protection in México, an estimated average of false calls from 2018 to 2024 is 77%. Likewise, its corresponding value from January to June 2024 is 73% (SSPC Executive Secretariat, 2024).

III. METHODOLOGY

Based on the identified processes from Step 2 identified as "Call Rings at PSAP" to Step 5 identified as "Units Notified / Dispatched," presented in Figure 1, for the call processing in a call center in a PSAT of an 911 Emergency Response System according to the National Emergency Number Association in the United States of America, we developed a stochastic simulation model to evaluate the *call answer interval* or *ringdown time.* Subsequently, we characterized the processes of call answering and dispatching based on probabilistic distributions of their corresponding times meeting or exceeding a p-value equal or larger than 0.05. Next, the arrival processes of every corporation were equally characterized probabilistically including police, medical and fire demand for service. Then, all characterizations of the call answering processes were configured in the stochastic simulation model.

Since data was not provided for false calls arrival pattern, an Exponential probabilistic distribution was utilized since it is commonly assumed that arrivals distribution in a service queue probably follows an Exponential distribution. For the duration of false calls we also assumed a mean of 1 minute, although shorter and larger durations may be observed if the answering agent and PSAP adhere to the protocols for managing false calls which may include a call back from the call center to the initiator of the call for service to evaluate if the person is under risk and if a unit dispatch is recommended (bNENA, 2020). Considering the volume of expected false calls of 73% in 2024, reported by the Executive Secretariat of the Secretary of Security and Citizen Protection (Secretariado Ejecutivo 911, 2024), we developed two model scenarios to manage false calls: (S1) without false calls, and (S2) with false calls. In scenario S2, we included a 70% volume of false calls.

Furthermore, for scenarios S1 y S2 we characterized performance of *ringdown time* for second level scenarios of one answering agent to eight answering agents to evaluate an ideal allocation of call answering agents that could generate an ideal ringdown time of 10, 15 or 20 seconds, as identified in the literature and established by Yung and Dayharsh (1980) and NENA (aNENA, 2020).

Our simulation models were run for 10 replicates of 552 hours and averages were considered to evaluate performance parameters.

IV. RESULTS

Inter arrival times were probabilistically characterized by type of corporation including police, medical, fire and two other corporations or classifications identified as Corporation 4 and Corporation 5. The inter arrival characterizations for the police corporation were conducted by patrolling zone and call priority for four quadrants integrated in one police district. Every police quadrant has four patrolling zones and calls for service have three levels of priority. A total of 48 probability characterizations were determined for the police corporation based on this configuration and results are shown in Table 1, which identifies the specific probability distributions found and their observed frequency for inter arrival and dispatch processes. Given a reduced volume of calls for service of the other remaining corporations compared with the volume of calls registered for the police related calls, they were characterized for the complete police district.

Table 1: Probability distributions for calls: *inter arrival time* and *answer + dispatch time*

Parameter	Number of Probability Distributions (95% C.I.)				Total
	Exponential	Gamma	Lognormal	Weibull	
*Inter arrival Time*s: *Police* Patrolling Zones: 4 Quadrants, 4 Patrolling Zones / Quadrant, 3 Call Priorities	1	21	14	12	48
Inter arrival Times: Medical *Fire* *Corporation 4* *Corporation 5*	1		1 1 1		1 1 1 1
Inter arrival Times: *False Calls*	1				1
Answer Call Time + Dispatch Time: *Police, Medical, Fire,* *Corporation 4* *Corporation 5* 4 Quadrants and 3 Call Priorities		2	10		12
Answer Call Time: *False Calls*	1				1
Total	4	23	27	12	66

Particularly, for the inter arrival times of police corporation, we observe that their probability distributions are represented by 2% Exponential, 43.7% Gamma, 29.1% Lognormal, and 25% Weibull. Similarly, the inter arrival times of the demand for service calls for Medical, Fire, Corporation 4, and Corporation 5 followed an Exponential for the Medical, and Lognormal distributions for Fire, Corporation 4 and Corporation 5. As stated before, the inter arrival times for False calls, was assumed to follow an Exponential probability distribution.

Service times for Call Answering and Dispatch are also presented in Table1. In this processes we identified for corporations of Police, Medical, Fire, Corporation 4, and Corporation 5 that their probability distributions were characterized by 16.6% Gamma, and 83.3% Lognormal, which agrees with Gualandi and Toscani (2019), who state that service times in call centers can be modeled by Lognormal probability distributions. Likewise with the Exponential probability distribution assumption for the inter arrival times of False calls, it was also assumed that the Call Answering process of False calls, also followed an Exponential probability distribution.

Our simulation results are presented in Figures 2 to 5. Figure 2 illustrates Scenario S1, where no false calls are included and the four police Quadrants (Q1 to Q4) of the evaluated police district represent the average of ten replicates of the waiting time or *ringdown time* that calls wait in the calls incoming quadrant locations before calls are answered, for every one of the eight second level scenarios to represent one to eight call answering agents. In this figure it is observed that if the call center only had one agent the expected *ringdown time* would be of 304 seconds to 395 seconds and if the call center had five agents the *ringdown time* would be of 19.3 seconds to 31.9 seconds. Evaluating this behavior, we look for a minimum number of agents that could generate a *ringdown time* equal or smaller than 20 seconds, 15 seconds, and 10 seconds, which are the minimum performance restrictions established by Yung and Dayharsh (1980) and NENA (aNENA, 2020). In this figure, the scenario with 8 servers or agents generates a *ringdown time* of 10.3 seconds to 16.7 seconds, which exceeds the 10 and 15 seconds restrictions, and meets the 20 seconds restriction. This type of analysis assist us in determining the minimum number of servers or agents that are able to meet the minimum performance parameters of *ringdown time*.

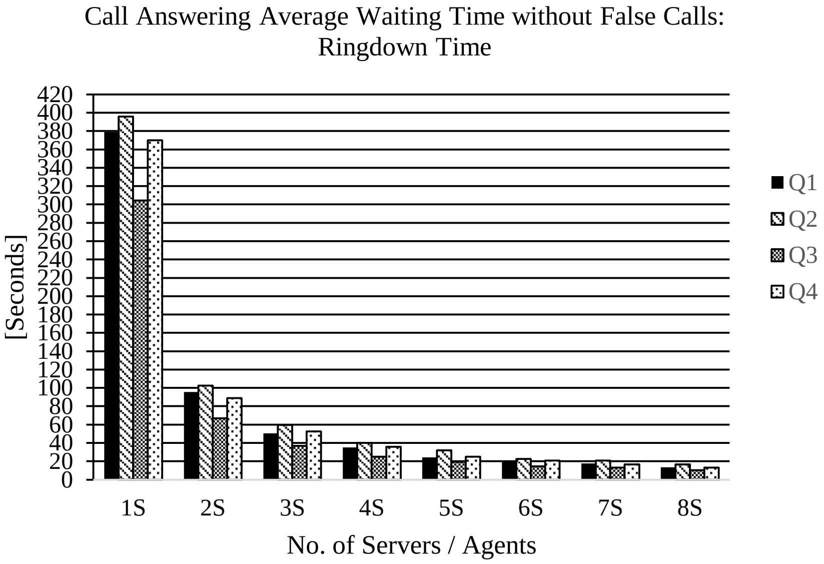

Figure 2: Police district quadrant (Q) averages of ten replicates: Estimated call answering waiting time (*ringdown time*) by number of servers without False Calls (Scenario S1)

Similarly to Figure 2, in Figure 3 we present results for scenario S2 for the *ringdown time* averages of the four police Quadrants (Q1 to Q4) of the evaluated police district of ten replicates, for every one of the eight second level scenarios to represent one to eight call answering agents, where False calls were included. Considering that: (1) False calls have a smaller time for the call answering process than legitimate calls, which was assumed to be probabilistically distributed based on an Exponential distribution with a mean of one minute, and (2) answering waiting time or *ringdown time* in our model is estimated based on average time a call stays in the waiting location before it is answered, and the average waiting time or *ringdown time* obtained in Scenario S2 is smaller than in Scenario S1. In Figure 4, the model of Scenario 2 with one server was utilized to evaluate call answering waiting time or *ringdown time* for False calls with answering time with a mean of 2 and 3 minutes with an Exponential distribution. As it can be observed in Figure 4, the *ringdown times* for Q1, Q2, and Q4 in scenario

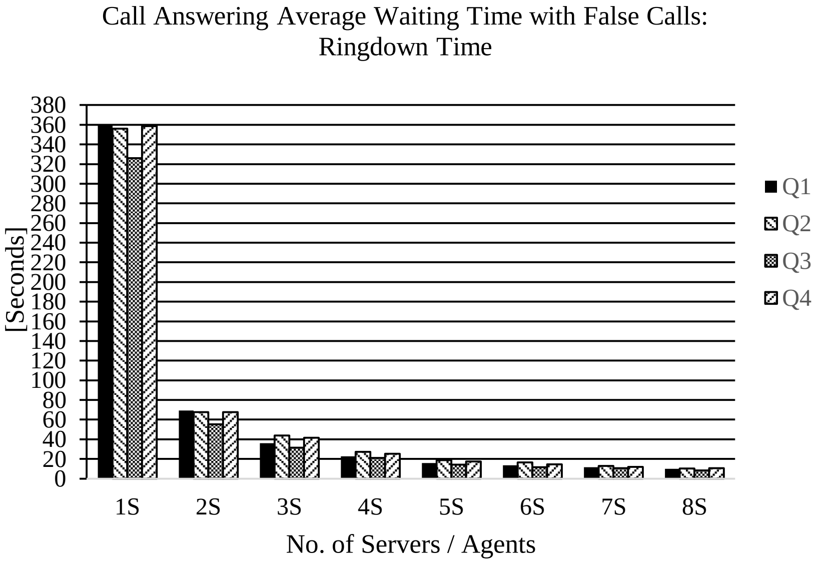

Figure 3: Police district quadrant (Q) averages of ten replicates: Estimated call answering waiting time (*ringdown time*) by number of servers with False Calls (Scenario S2)

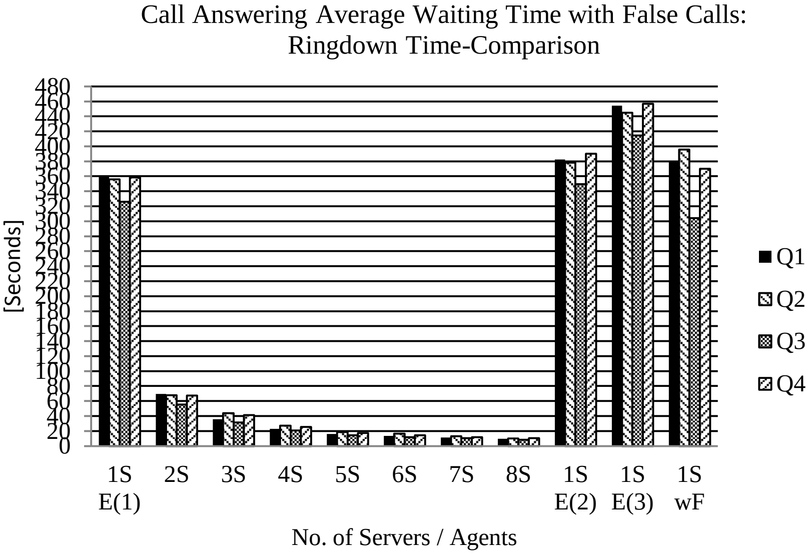

Figure 4: Police district quadrant (Q) averages of ten replicates: Estimated call answering waiting time (*ringdown time*) by number of servers with False Calls – Comparison

with one server without False calls (1S wF) are larger than Q1, Q2, and Q4 of the same scenario with 1 server but with false calls (1S E(1)). However, when call answering time for False calls is changed from Exponentially distributed with mean of 1 minute to Exponentially distributed with mean of 2 (E2) and 3 minutes (E3), the call answering waiting time or ringdown time is increased and averages in 1S E(2) and 1S E(3) are larger than averages in 1S wF except for Q2 in 1S E(2).

In Figures 5 and 6, we present the maximum contents of calls in incoming locations queues without and with False calls corresponding to Scenario 1 and Scenario 2 respectively. As illustrated in Figure 5, the average maximum contents of calls is presented when the system has only one server or agent reaching a low value of 13 calls and a high value of 13.7 calls. Similarly, when the system has eight servers or agents the maximum content of calls has a low value of 10.8 calls and a high value of 11.2 calls. These values are the averages of the maximum contents of calls during the simulation run and are not average values of the average values given the ten simulation replicates.

Likewise, in Figure 6 that presents scenario S2 including False calls, we also observe the maximum content of calls when the system has only one server or agent with a low value of 15.2 calls and a high value of 16.1 calls. Also, the system presents the smallest maximum values when the system has eight servers or agents with a low value of 11.9 calls and a high value of 12.3 calls.

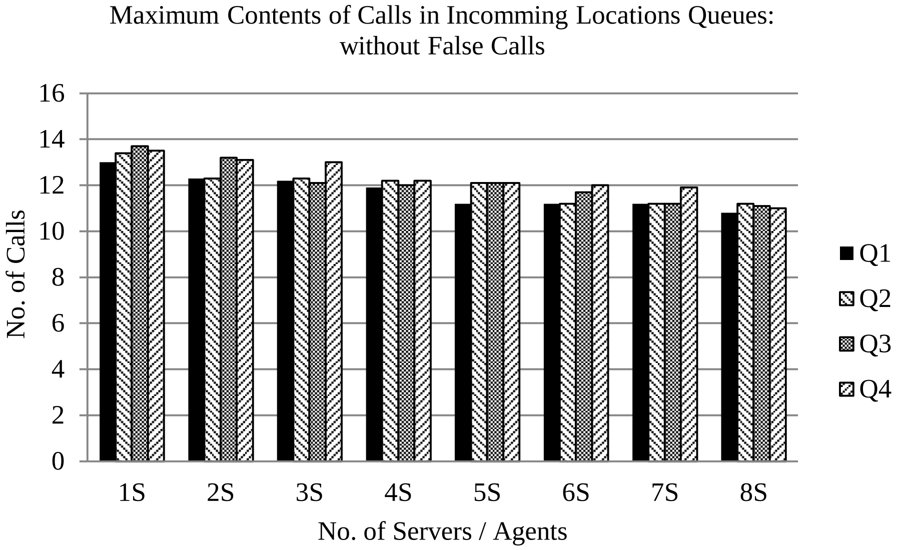

Figure 5: Police district quadrant averages of ten replicates: Maximum content of calls in the calls receiving locations by number of servers or agents (Scenario 1)

From Figures 5 and 6 we observe that the volume of False calls does impact in the maximum content of calls which may be observed in any given time.

V. DISCUSSION AND CONCLUSION

Although performance parameters such *response time* and *ringdown time* of Emergency Response Systems (ERS) and their Call Centers are not easily accessible, through the application of analytic tools such as discrete event stochastic simulation and limited data of the ERS, analysis and improvement of their services could be possible. In our case, we were able to indirectly estimate *response time* and *ringdown time* of present operating strategies of the ERS and its Call Center by applying the tool of discrete event stochastic simulation. At the same time, the tool of discrete event stochastic simulation allowed us to model plausible scenarios to evaluate potential performance improvements. The characterization of the level of resource allocation such as the evaluation of gradually increasing the number of call answering servers or agents was possible, and very useful to determine with precision, the minimum or ideal number of servers or agents to the ERS Call Center to be able to meet, in this case, the maximum *ringdown times* established by Yung and Dayharsh (1980) and NENA (aNENA, 2020).

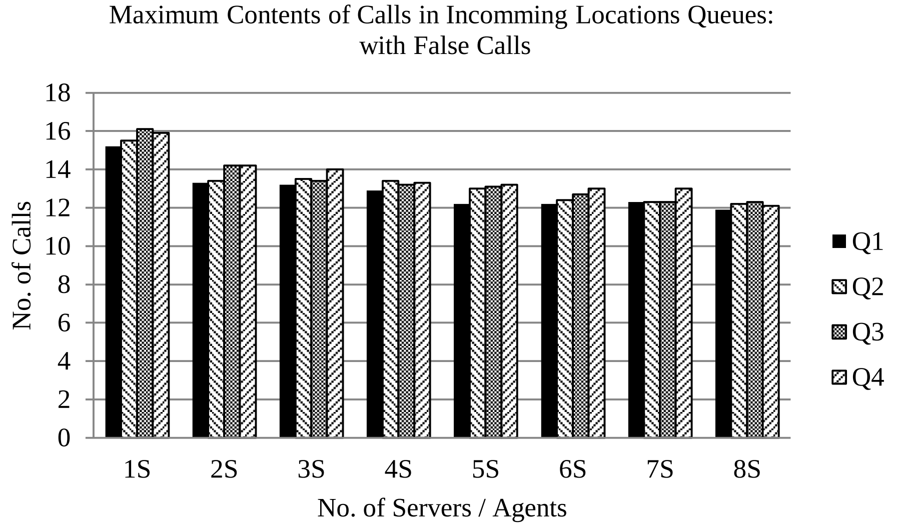

Figure 6: Police district quadrant averages of ten replicates: Maximum content of calls in the calls receiving locations by number of servers or agents (Scenario 2)

We believe that in order to achieve performance goals such as ideal *response time* and *ringdown times* in an ERS and its Call Center, there must be mandatory policies or guidelines and allocate a required budget. In parallel, it is suggested to establish a continuous supervision and improvement of the ERS system.

REFERENCES

[1]. Brooks, J.P., Edwards, D.J., Sorrel, T.P., Srinivasan, S. and Diehl, R.L. 2011. Simulating Calls for Service for an Urban Police Department. *Proceedings of the 2011 Winter Simulation Conference,* Phoenix, AZ, USA, December 11, 2011.

[2]. Chanpanit, T. and Udomsakdigool, A. 2020. Big Data Framework for Incoming Calls Forecasting in a Call Center. *Proceedings of the 2nd International Conference on Electrical, Communication and Computer Engineering (ICECCE),* Istanbul, Turkey, June 12, 2020.

[3]. Conley, Q. and Grabau, M. 2013. Simulating Modified Hybrid Approach to Resource Assignment in a Shared Billing and Claims Call Center. *Proceedings of the 2013 Winter Simulation Conference*, Washington, D.C., USA, December 8, 2013.

[4]. D'Amico,S.J., Wang, S., Batta, R., and Rump, C.M. 2002. A Simulated Annealing Approach to Police District Design. Computers and Operations Research, 29, 667-684.

[5]. Gualandi, S. and Toscani, G. 2019. Human Behavior and Lognormal Distribution: A Kinetic Description. Mathematical Models and Methods in Applied Sciences, 29:4, 717-753.

[6]. Hatami-Marbini, A., Varzgani, N., Sajadi, S.M., Kamali, A. 2022. An emergency medical services system design using mathematical modeling and simulation-based optimization approaches. Decision Analytics Journal, https://doi.org/10.1016/j.dajour.2022.100059

[7]. Huang, Y. 2015. Modeling and Simulation Method of the Emergency Response Systems based on OODA. Knowledge-Based Systems, 89, 527-540

[8]. INEGI 2023. Encuesta Nacional de Victimización y Percepción sobre Seguridad Pública ENVIPE 2023. Instituto Nacional de Estadística Geografía e Informática. Aguascalientes, Ags., México.

[9]. Ibrahim, R., L'Ecuyer, P., Regnard, N. and Shen, H. 2012. On the Modeling and Forecasting of Call Center Arrivals. *Proceedings of the 2012. Winter Simulation Conference,* Berlin, Germany, December 9, 2012.

[10]. Jennex, M.E. 2007. Modeling Emergency Response Systems. *Proceedings of the 40th Hawaii International Conference on System Sciences – 2007*. Big Island, Hawaii, Jan. 3 2007 to Jan. 6 2007.

[11]. L'Ecuyer, P., Gustavsson, K., and Olsson, L. 2018. Modeling Bursts in the Arrival Process to an Emergency Call Center. *Proceedings of the 2018 Winter Simulation Conference.* Gothenburg, Sweden, December 2018.

[12]. Li, S., Koole, G. and Jouini, O. 2019. A Simple Solution for Optimizing Weekly Agent Scheduling in a Multi-Skill Multi-Channel Contact Center. *Proceedings of the 2019 Winter Simulation Conference, National Harbor, MD, U.S.A., December 8, 2019..*

[13]. Liao, S., Van Delft, Ch., Koole, G., Dallery, Y. and Jouini, O. 2009. Call Center Capacity Allocation with Random Workload. 2009 International Conference on Computers & Industrial Engineering, Troyes, France, July 6, 2009.

[14]. (a) National Emergency Number Association (NENA) PSAP Operations Committee, 9-1-1 Call Processing Working Group. 2020. NENA Standard for 9-1-1 Call Processing. NENA, Alexandria, VA, U.S.A.

[15]. (b) National Emergency Number Association (NENA). 2020. A Study Focused on Processing Silent or Hang-Up 9-1-1 Calls for Service. NENA, Alexandria, VA, U.S.A.

[16]. Piyadasum, T., Kalansuriya, B., Gangananda, M., Malshan, M.,Bandara, D.H.M.N., Marru, S. 2017. Rationalizing Police Patrol Beats Using Heuristic-Based Clustering. *Moratuwa Engineering Research Conference (MERCon), Moratuwa, Sri Lanka, May 29, 2017.*

[17]. Seada, A.A. and Eltawil, A.B. 2015. Modeling and Analysis of Workforce Management Decisions in Modern Call Centers. *Proceedings of the 2015 International Conference on Industrial Engineering and operations Management, Dubai, United Arab Emirates (UAE), March 3, 2015.*

[18]. Secretariado Ejecutivo 911. 2024. Estadística nacional de llamadas de emergencia al número único 9-1-1Cifras con corte al 30 de junio de 2024. Centro Nacional de Información (CNI) Julio 2024.Secretaría de Seguridad y Protección Ciudadana, Cd. de México, México.

[19]. Steinmann, G. and de Freitas Filho, P.J. 2013. Using Simulation to Evaluate Call Forecasting Algorithms for Inbound Call Center. *Proceedings of the 2013 Winter Simulation Conference, Washington, D.C., USA, December 8, 2013*.

[20]. Ta, T. A., Chan, W., Bastin, F., L'Ecuyer, P. 2021. A simulation-based decomposition approach for two-stage staffing optimization in call centers under arrival rate uncertainty. European Journal of Operational Research, 293:3, 966-979.

[21]. van Barneveld, T., Jagtenberga, C., Bhulaia, S., and van der Mei, R. 2018. Real-Time Ambulance Relocation: Assessing Real-Time Redeployment Strategies for Ambulance Relocation. Socio-Economic Planning Sciences, 62: 129-142.

[22]. van Buuren, M., Kommer, G.J., van der Mei, R. and Bhulai, S. 2015. A Simulation Model for Emergency Medical Services Call Centers. *Proceedings of the 2015 Winter Simulation Conference,* Huntington Beach, CA, U.S.A.

[23]. van Buuren, M., Kommer, G.J., van der Mei, R. and Bhulai, S. 2017. EMS call center models with and without function differentiation: a comparison. Operations Research for Health Care, 12: 16-28.

[24]. Ye, X., Chen, B., Lee, K., Storesund, R., Li, P., Kang, Q, Zhang, B. 2021. An emergency response system by dynamic simulation and enhanced particle swarm optimization and application for a marine oil spill accident. Journal of Cleaner Production, 297: https://doi.org/10.1016/j.jclepro.2021.126591

[25]. Yu, M., Zhang, M. and Jiao, H. 2018. Capacity Sizing in the Presence of Repeated Customer Behavior: Dimensioning an Inbound Call Center. *2018 Chinese Control and Decision Conference (CCDC), Shenyang, China, June 9, 2018.*

[26]. Yung, T.J. and Dayharsh, T.I. 1980. The Design and Costing of 911 Systems- -A Technical Manual. U.S. Department of Justice, Bureau of Justice Statistics, Washington, D.C.