

Article

Simulated data to estimate real sensor events: A Poisson-regression-based modelling

Miguel Ortíz-Barrios ^{1,*} , Ian Cleland ^{2,**} , Chris Nugent ^{2,***} , Pablo Pancardo ^{3,†} , Eric Järpe ^{4,††}  and Jonathan Synnott ^{2,‡} 

¹ Department of Industrial Management, Agroindustry and Operations, Universidad de la Costa CUC, Barranquilla, Colombia

² School of Computing, Computer Science Research Institute, Ulster University, Belfast, BT37 0QB, UK

³ Academic Division of Information Science and Technology, Juarez Autonomous University of Tabasco, Tabasco, Mexico

⁴ Department of Intelligent Systems and Digital Design, Halmstad University, Halmstad, Sweden,

* mortiz1@cuc.edu.co

** i.cleland@ulster.ac.uk

*** cd.nugent@ulster.ac.uk

† pablo.pancardo@ujat.mx

†† eric.jarpe@hh.se

‡ j.synnott@ulster.ac.uk

* Correspondence: mortiz1@cuc.edu.co

Version February 15, 2020 submitted to *Remote Sensing*

Abstract: Automatic detection and recognition of **Activities of Daily Living (ADL)** are crucial for providing effective care to frail older adults living alone. A step forward in addressing this challenge is the deployment of smart home sensors capturing the intrinsic nature of ADLs performed by these people. As the real-life scenario is characterized by a comprehensive range of ADLs and smart home layouts, deviations are expected in the number of sensor events per activity (SEPA), a variable often used for training activity recognition models. Such models, however, rely on the availability of suitable and representative data collection and is habitually expensive and resource-intensive. Simulation tools are an alternative for tackling these barriers; nonetheless, an ongoing challenge is their ability to generate synthetic data representing the real SEPA. Hence, this paper proposes the use of Poisson regression modelling for transforming simulated data in a better approximation of real SEPA. First, synthetic and real data were compared to verify the equivalence hypothesis. Then, several Poisson regression models were formulated for estimating real SEPA using simulated data. The outcomes revealed that real SEPA can be better approximated ($R^2_{\text{pred}} = 92.72\%$) if synthetic data is post-processed through Poisson regression incorporating dummy variables.

Keywords: Activity recognition; Activities of Daily Living (ADL); Digital Simulation; Poisson regression; Large-scale datasets; Sensor systems; Smart homes

1. Introduction

Remote sensing is enabling us to understand more about our surroundings, particularly around environmental change. Remote sensing through geospatial data, is however, not typically seen as a means for continuous monitoring. It generally relies on sensors attached to aircraft or satellite for geological mapping or capturing observations of the earth. As a result, remote sensing is often associated with collection frequencies measured in months rather than hours or days. There are a vast number of monitoring and inspection applications that would require and benefit from more frequent observation. For these applications, remote sensing and Internet of Things (IoT) could be used to

25 complement and strengthen each other. Remote sensing and IoT bring together external observations
26 possible only from extrinsic sensors such as satellite images and combine/rationalize these findings
27 with data streamed by embedded IoT sensors.

28 Convergence of these two distinct sensor technologies is being driven by the emergence of
29 platforms, such as robotics, autonomous drones and spectrographic sensors, which can be mounted on
30 these smaller platforms. The mobility and costs of these new platforms make them ideal for constant
31 deployment rather.

32 Advances in the disciplines of remote sensing and IoT are blurring the distance between the two
33 previously distinct worlds. Both movements are motivated by similar needs, however, have evolved
34 based on the circumstances under which each was conceived. For example, both emerged from the
35 need to collect data efficiently and at scale without requiring humans to create the data. Both have also
36 evolved to create data and analytics that reduce vast amounts of data into actionable insights through
37 inference. This in turn requires large amounts of representative data.

38 This paper discusses how IoT data can be synthetically generated in a robust and representative
39 manner, in order to develop and test machine intelligence models for data processing and inference.
40 Whilst this method is demonstrated within a smart home use case, the results are also applicable to
41 other application areas including industry 4.0, energy management and transformational health and
42 transportation services. Synthetic data generated by simulation are useful to complement real data in
43 an IoT environment so that human behaviour may be detected and modeled [1,2]. If machine/deep
44 learning algorithms are also applied, it is possible to have solutions for activity recognition, fall
45 detection, behaviour modelling and risk determination [3,4].

46 In particular, ageing of the world population and the impossibility of having relatives or caregivers
47 to take care of them at all times brings the need for remote surveillance, Activity Daily Living (ADL)
48 assistance, accompaniment, support for medication, among others. In this sense, researchers have
49 developed proposals that contribute to alleviating the situation, suggestions are varied and include
50 intelligent solutions to assist the elderly and enable them to preserve or improve their quality of life.
51 However, the continued development and improvement of these solutions rely on the availability of
52 large amounts of accurately labelled and representative data in order to train and evaluate developed
53 techniques [5,6]. Limitations to the gathering and availability of ADL data have previously been
54 highlighted as detrimental to research progress, potentially slowing advances in the field [7–9].
55 Moreover, data acquisition is often not feasible due to the complexity of having older adults willing to
56 stay in a laboratory and perform natural daily activities; besides, the time needed is also ample.

57 Much effort has been made to produce fully annotated and publicly available datasets for
58 benchmarking, for example [10–12]. These datasets are, however, usually limited in terms of the
59 number of inhabitants, the range of activities considered, the types of sensors deployed and the size
60 and layout of the environment. In many cases, datasets focus on a single occupant recorded in one
61 small apartment. These limitations can be attributed to the prohibitive costs associated with both
62 creating a real smart environment and the time required for data collection and annotation [13,14].
63 This makes having large scale deployments across multiple environments even more challenging.

64 With this in mind, many researchers have been looking for alternative methods of data collection,
65 this includes standard protocols for data sharing [15], model-based synthetic data generation and
66 interactive data simulation tools [16]. In particular, simulation tools represent a very promising
67 alternative to real environments as they provide a flexible and cost-effective solution for the collection
68 of realistic data [13].

69 Consequently, simulated data to complement real data is a common solution. However, the
70 problem is that data from simulation tools, in some cases, is not reflective of that collected within
71 a real environment, so, our solution proposal consists of generating some simulated data for each
72 of the possible events in the activities considered, as well as the different types of sensors involved.
73 Later, we analyze simulated data to identify those that are statistically different to real data, then we
74 transform them applying a mathematical method, in such a way that data generated with simulation

75 can effectively complement real data. So that more significant volumes of data can be achieved for
76 training and calibration of activity classification models.

77 In the literature, many research works make use of the combination of real data and simulated
78 data for the feeding of ADL classification models. For example, in [17] was implemented a smart-home
79 simulation combining an avatar-based scenario (acquired from real-world data), and probabilistic
80 **modelling** of sensors. Authors obtained similarities of simulated and real data, so it demonstrated the
81 viability of probabilistic sampling approach.

82 The proposal exposed in this article is an extension of a previous work [18] in which we applied
83 different regression models to simulated data to use them as a complement to real data. In our previous
84 work, a simulation was carried out in terms of duration and intensity of sensor events; however, we
85 did not consider different types of sensors that **associate to** each ADL and that are involved in events
86 that form **activity sequences**. **Precisely, the simulation considering the different types of sensors are**
87 **studied in the paper and applying Poisson regression to improve the results. The main difference to**
88 **the previous work is that here we hypothesize that data generated by simulation and adjusted by**
89 **Poisson regression is more similar to real data than unadjusted data.**

90 This paper is described as follows: Section 2 presents a review of simulation tools that have
91 been designed for smart environments; whereas Section 3 depicts the details on Poisson regression
92 modelling. **Section 4 describes our experiments whereas** results are reported and discussed in Section
93 5. Finally, Section 6 details the conclusions and future work.

94 2. Review of Simulation Tools for Smart Environments

95 A comprehensive review of simulation tools has been reported previously within the literature.
96 [13,16] These tools can be split into **model-based** and interactive approaches. **Model-based** approaches
97 generally focus on the use of statistical or machine learning techniques to generate synthesised or
98 surrogate data [19,20]. **Techniques involved in this include the use of correlation preservation, and**
99 **amplitude distribution [21] or more recently the use of adversarial neural networks [22].** This section
100 will provide an overview of interactive approaches and approaches used to validate the generated
101 data.

102 Generally, interactive approaches rely on a human user controlling an avatar around a 2D or 3D
103 virtual Environment [16]. As the avatar moves throughout the environment, it interacts with various
104 passive and active virtual sensors and/or actuators, for example activating pressure or presence
105 sensors and turning on or off lights.

106 The intelligent environment simulation (IE Sim), developed by Synnott et al. [23] is a tool to
107 generate simulated datasets for Activities of Daily Living. It allows the researcher to design smart
108 homes by providing a 2D graphical top-view of the floor plan. The researcher can add different types
109 of sensors such as temperature sensors and pressure sensors. Using an avatar, the researcher can
110 carry out ADLs interacting with objects and triggering sensor in the virtual environment. Similarly,
111 Ariani et al. [24] created a smart home simulation tool that collects data from virtual ambient sensors
112 including binary motion detectors and pressure sensors. The researcher produces the smart home floor
113 plan by drawing shapes on a 2D canvas and can then place sensors onto the floor plan. To simulate
114 the activities and interactions in the smart home, the authors used a pathfinding algorithm which
115 simulates the movement of the inhabitants.

116 **The OpenSHS simulator [13], generates realistic Smart Home data through a hybrid approach.**
117 **Specifically, it combines both interactive and model-based approaches. Data generated through**
118 **interactive simulation can then be replicated using a specifically designed algorithm. The OpenSHS**
119 **was demonstrated in generating a dataset for classification as well as the detection as anomalous**
120 **activity such as leaving the front door open. The opensource simulator, SIMACT [25] allows for the**
121 **creation of a 3D environment and the selection and positioning of virtual sensors. These virtual sensors**
122 **are modelled upon common Smart Home sensors such as RFID, PIR sensors, and contact sensors. The**
123 **simulator generates datasets in two modes 1) is an interactive mode where the avatar is controlled by**

124 the user who can interact with various items in the home, 2) is a model-based approach where the
 125 inhabitants are controlled by pre-defined scripts where the user defines the completion time of each
 126 step and the objects that are interacted with.

127 As highlighted by Table 1, few of the interactive approaches reported in the literature have
 128 compared the accuracy of data generated by the simulator with data generated in a real environment.
 129 When doing so, it is important to consider not only which sensors are firing, but also the duration and
 130 timing of these sensor events. For instance, making a meal may take a longer time in the morning than
 131 in the evening.

<i>Author</i>	<i>Date</i>	<i>2D/3D</i>	<i>Comparison with real data</i>
OpenSHS [13]	2017	3D	No
Park [26]	2015	3D	No
PerSim [27]	2015	3D	Yes
IE Sim [23]	2015	3D	Yes
Ariani [24]	2013	2D	No
SimCon [28]	2010	3D	No

Table 1. Table summarizing notable works which have produced interactive simulation tools for Smart environments. This highlights the lack of comparison with real world datasets.

132 Lee et al. [27] developed the PerSim 3D human activity simulator. A contrast between real data
 133 gathered within the Gator Tech Smart House and synthetic data produced by Persim 3D concluded
 134 mean data similarities of between 0.78 and 0.81. Another work comparing real data and with data
 135 produced by the simulator MASSHA25 revealed the similarities between 0.8810 and 0.9352 in terms of
 136 frequency, and 0.9827 and 0.9909 in terms of duration on datasets including single user ADLs.

137 Renoux and Klugl [29] presented a framework to generate sensor data from a simulated smart
 138 home. The solution used a flexible agent-based simulation tool and constraint-based planning. The
 139 authors highlighted that the data generated could be used to test or train algorithms that are then
 140 directly usable in real-world applications. Through an evaluation of the solution, the authors showed
 141 that the activity plans generated by the simulator show some plausibility. The comparison of these
 142 plans with real datasets, however, revealed some issues. In particular, there was a noted discrepancy
 143 between the expectation of what an activity plan for a real day looked like when looking at a complete
 144 day and the actual recorded timeline of a real day. The author was unclear whether this difference was
 145 due to real activity fragmentation or to errors during activities annotation.

146 As discussed, interactive methods mainly depending on an avatar interacting within a virtual
 147 environment to produce simulated datasets have a restricted capability to consider the inherent
 148 variations in activity duration and intensity regarding the daytime that would be exhibited in
 149 a real dataset. This is mainly due to the synthetic nature through which interaction takes place.
 150 Mendez-Vazquez et al. [30] demonstrated the use of Markov chains describing the order of events,
 151 combined with Poisson distribution to calculate a range of realistic activity times and probability
 152 distributions to calculate a range of sensor values to generate a simulated activity dataset. This
 153 simulated activity set contained a distribution of activities such as reading, sleeping, walking and
 154 sitting together with metrics including time and energy expenditure. As a result, simulated datasets
 155 may not be reflective of those produced in a real environment.

156

157 3. Method

158 The procedure of Poisson regression (see e.g. [31]) is a generalized linear model where the log
 159 conditional expected response given the covariates can be expressed as a linear combination of the
 160 covariates and a noise term, i.e.

$$\log E(Y|X) = \beta_0 + \sum_{k=1}^n \beta_k X_k + \epsilon \quad (1)$$

161

162 where Y is the response variable, X_1, X_2, \dots, X_n are the covariates and ϵ is the noise term. The response
 163 is a variable of count data and is assumed to be Poisson distributed. **Poisson regression is adopted**
 164 **in this study since it is suitable for modelling the relationship between a group of predictors (in this**
 165 **case: simulated events per activity, simulated activity duration, and simulated events per sensor) and**
 166 **a response variable representing the number of times an event occurs in a finite timestamp (as real**
 167 **SEPA does).**

168

169 Other candidates to modelling the activity durations could be e.g. ordinary regression with
 170 Gaussian response, non-linear regression models, models for data subrogation [32], non-Gaussian
 171 models etc. However, the Poisson distribution is fundamentally derived from the assumption of
 172 counts. More to the point, when there is no reason to assume dependence between events, resulting
 173 in either clustering or regularity, the number of events in each fixed interval is Poisson distributed.
 174 Here, the covariates are typically count data (number of events per activity and number of events per
 175 sensor), but also time (duration of an activity). The linear combination of these counts (possibly also
 176 including interaction terms) is therefore hypothesized to be Poisson rather than Gaussian or otherwise
 177 distributed. Gaussian and other continuous distributions are preferable when the data is continuous
 (such as exact weight, length etc). See [33] for further motivations for choice of model.

178 3.1. Overdispersion

179

The (homogeneous intensity) Poisson probability function of a single variable Y is

$$P(Y = y) = f(y; \lambda) = \frac{\lambda^y e^{-\lambda}}{y!} \quad (2)$$

180

181 for $y \in \{0, 1, 2, \dots\}$ and $\lambda \in \mathbb{R}^+$. A characterizing property of the Poisson distribution is that the
 182 single parameter stands for both expectation and variance. If these two moments differ, this is an
 183 indication that the Poisson assumption is not sustained. The phenomenon of the variance being larger
 184 than the expectation is called *overdispersion*, typically caused by too few covariate observations or
 185 strongly correlated covariates. In case when the dispersion is on par with the expectation the term is
 186 *equidispersion*.

186

To address the problem of overdispersion one may consider the Poisson distribution as a special
 187 case of the Generalized Poisson (GP) distribution, with a probability function

$$P(Y = y) = f(y; \lambda, \kappa) = \frac{\lambda(\lambda + \kappa y)^{y-1} e^{-\lambda - \kappa y}}{y!} \quad (3)$$

188

189 for $y \in \{0, 1, 2, \dots\}$, $\lambda \in \mathbb{R}^+$ and $\max(-\frac{\lambda}{4}, -1) < \kappa < 1$. Then the case of Poisson distribution
 190 corresponds to GP distribution with $\kappa = 0$ (equidispersion), overdispersion corresponds to $\kappa > 0$
 191 and underdispersion corresponds to $\kappa < 0$. A random variable X distributed according to the GP
 distribution has expectation and variance (see [34]) according to

$$E(Y) = \frac{\lambda}{1 - \kappa} \quad V(Y) = \frac{\lambda}{(1 - \kappa)^3} \quad (4)$$

192

193 which motivates defining the dispersion parameter ϕ as the deviance of the GP from the Poisson
 distribution with $\phi = \frac{1}{(1 - \kappa)^2}$.

194

195 A hypothesis test for checking on indications of overdispersion may be carried out by utilizing
 the Pearson's goodness-of-fit statistic

$$X^2 = \sum_{i=1}^n \frac{(Y_i - \hat{\mu}_i)^2}{\hat{\phi} \hat{\mu}_i} \quad (5)$$

196 which is χ^2 -distributed with $n - 1$ degrees of freedom under the null hypothesis $H_0 : \kappa = 0$ against
 197 the alternative $H_1 : \kappa > 0$. Here, x_{ij} denotes the i th observation of the j th covariate X_j , the estimator
 198 $\hat{\mu}_j = \sum_{j=1}^m \hat{\beta}_j x_{ij}$ and the dispersion parameter estimator $\hat{\phi}$ is achieved by maximum likelihood
 199 estimation in parallel to the estimation of the regression coefficients [35].

200

201 3.2. Normally distributed residuals

202 One assumption in the **modelling** with Poisson regression is that the residuals **follow a normal**
 203 **distribution**. To validate this condition an Anderson-Darling test **is conducted to confirm** that the
 204 residuals are not deviating significantly from the normal distribution. Assuming the expected value
 205 of the residuals $E(\varepsilon) = \mu_\varepsilon = 0$, the standardized residuals are $\tilde{\varepsilon}_i = \varepsilon_i \sqrt{n} (\sum_{i=1}^n \varepsilon_i^2)^{-1/2}$ and the
 206 Anderson-Darling test statistic is

$$A^2 = (25n^{-3} - 4n^{-2} + 24n^{-1} - 4 - n) \sum_{i=1}^n (2i - 1) \log [\Phi(\tilde{\varepsilon}_i)(1 - \Phi(\tilde{\varepsilon}_{n-i+1}))]. \quad (6)$$

207 where $\Phi(\cdot)$ denotes the standard normal cdf. This statistic may then be used to reject the null
 208 hypothesis that the residuals are normally distributed in favour of the alternative that the residuals are
 209 not normally distributed as soon as the value of A^2 exceeds the Anderson-Darling percentile [36].

210

211 3.3. Independence

212 Another assumption is that the covariates are uncorrelated within each sample, i.e. that the
 213 auto-correlation $\rho(h) = \frac{C(X_i, X_{i+h})}{\sqrt{V(X_i)V(X_{i+h})}} = 0$ for all $h \neq 0$. This may be estimated by

$$r(h) = \frac{1}{n-h} \sum_{i=1}^{n-h} (x_i - \bar{x})(x_{i+h} - \bar{x}) \quad (7)$$

214 for time lags $h = 0, 1, 2, \dots, n - 1$ where \bar{x} is the average $\frac{1}{n} \sum_{i=1}^n x_i$. To check whether there is evidence
 215 for dependence violating the assumptions, one may perform an ordinary t -test of whether or not
 216 $\rho(h_0) \neq 0$ for some specified lag h_0 . Another way is to check the evidence for dependence is to utilize
 217 a Ljung-Box test to multiply check whether correlations $\rho(h) \neq 0$ for all lags h such that $|h| \leq h_0$ for
 218 some specified bound h_0 [37]. This may be carried out by calculating the test statistic

$$Q = n(n - h_0) \sum_{h=1}^{h_0} \frac{r(h)^2}{n - h} \quad (8)$$

219 and reject the null $H_0 : [\rho(1) = 0, \rho(2) = 0, \dots, \rho(h_0 - 1) = 0 \text{ and } \rho(h_0) = 0]$ in favour of the alternative
 220 $H_1 : [\rho(1) \neq 0, \rho(2) \neq 0, \dots, \rho(h_0 - 1) \neq 0 \text{ or } \rho(h_0) \neq 0]$ at level α of significance as soon as Q exceeds
 221 $\chi_\alpha^2(h_0)$ (the chi-square α -percentile with h_0 degrees of freedom).

222 4. Experiment description

223 Smart environments were developed to help older **people** or people who are suffering from
 224 some degenerative disorders (i.e. dementia) to maintain their independence in daily life. This is the
 225 case of the Halmstad Intelligent Home (HINT) at Halmstad University (Sweden), where a realistic
 226 home environment is provided for underpinning innovations and research studies relating to human
 227 behaviour analysis [38]. HINT, an apartment of $50m^2$ built, is supplied with a variety of thermal
 228 cameras and sensors (PIR, pressure, door contact, contact/touch, and others) capable with supporting

229 i) emergency detection and on-time response, ii) detection of deviating behaviour patterns, and
 230 iii) healthcare monitoring [38]. The left and right sides of Fig.1 present some of the spaces within
 231 this home lab. Such an environment was designed in IE Sim software which virtually incorporated
 232 the current sensor deployment so that model robustness and reliability can be fully granted. To
 233 compare real and **synthetic SEPA**, an experiment involving eleven participants was undertaken. Each
 234 participant was initially asked to carry out a set of eight ADLs (Go to bed, Use bathroom, Prepare
 235 breakfast, Leave house, Get cold drink, Go to office, Get hot drink, and Prepare dinner) in the virtual
 236 **environment** by using a virtual avatar (see Fig. 2).

237



Figure 1. Left: The HINT kitchen. Right: The dining and living rooms at Halmstad Intelligent Home

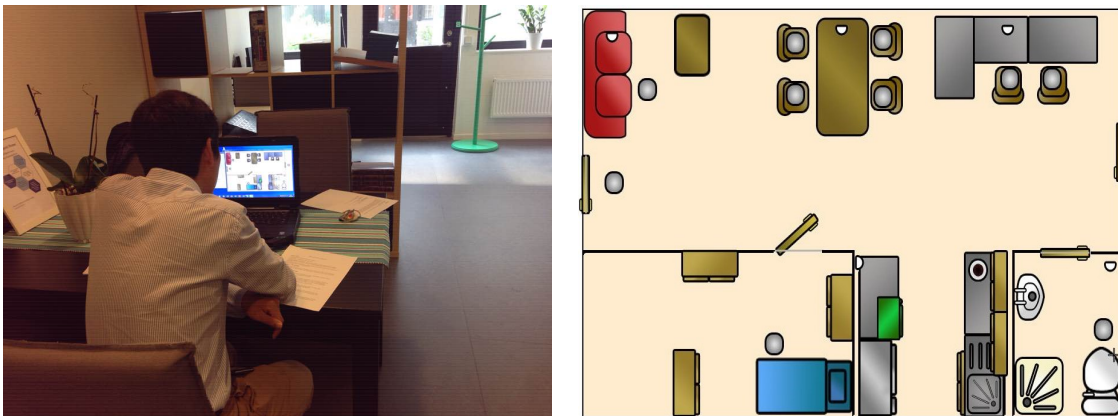


Figure 2. Left: User during the first simulation test with IE Sim. Right: The virtual representation of HINT environment designed within IE Sim.

238 A general description of the activities to be performed by users can be found below:

239 *Initial instructions*

240

- 241
- Please close each door after passing through.
 - 242
 - 243 • Please **turn off each domestic appliance after use.**
 - 244
 - 245 • You will be guided through each activity in sequence, please remember to select the
 - 246 "Stop/Start" button after each activity is complete.
 - 247
 - 248 • Time is not an issue in this experiment. Do not worry about needing to take time to re-read
 - 249 an activity description.

250

251 *Activity 1: Go to bed*

252 You can stay in bed all **the** time that you want. Time maximum is 2 minutes. Then, you have to
 253 leave the bedroom, close the door and press the button.

254
 255 *Activity 2: Use bathroom*

256 You can use **the** toilet if you need, or just wash **your** hands. Then, leave **the** bathroom, close **the**
 257 door, and press the button.

258
 259 *Activity 3: Prepare breakfast*

260 You have to prepare something to eat for breakfast. You can choose between milk and cereals or
 261 coffee, but you can make or also prepare both. Then, put the bowl on the table, sitting, and press
 262 the button.

263
 264 *Activity 4: Leave house*

265 You can choose to leave the home either from the front door or from the garden door. When you
 266 are outside, press the button.

267
 268 *Activity 5: Get cold drink*

269 You can choose between tap water or by taking something from the fridge. Then, put the glass
 270 with **the** drink on the kitchen desk and push the button.

271
 272 *Activity 6: Go to Office*

273 **You have to go to the office and press the button.**

274
 275 *Activity 7: Get hot drink*

276 You can choose between making tea or coffee. Then, put the cup on the kitchen desk and press
 277 the button.

278
 279 *Activity 8: Prepare dinner*

280 You have to prepare a soup. Put the bowl on the table and press the button.
 281

282 Once participants finished the aforementioned activities within the simulator, they were required
 283 to undertake the ADLs at HINT. The resulting data from real home and virtual environment were then
 284 arranged as two datasets specifying each sensor events aligned with their corresponding participant,
 285 sensor ID, code, sensor type, and time stamp. The next chapter will present the comparison between
 286 data emanating from HINT and IE Sim in terms of real SEPA. Furthermore, it will illustrate how
 287 synthetic data can be transformed (using Poisson regression modelling) for approximating the number
 288 of events perceived by each sensor in the real environment.

289 **5. Results and Discussion**

290 *5.1. Contrast between simulated and real sensor events*

291 **Paired t-tests ($\alpha = 5\%$, $CL = 95\%$) were performed to contrast the number of synthetic and real**
 292 **SEPA** considering two sensor types: door and pressure. This study also regarded seven ADLs (*Go to bed,*
 293 *Use bathroom, Prepare breakfast, Leave house, Get cold drink, Be in the office, and Prepare dinner*) and eleven
 294 sensors (*Bedroom door, Bed pressure, Bathroom door, Bowl cupboard–Prepare breakfast, Refrigerator–Prepare*
 295 *breakfast, Chair pressure–Prepare breakfast, Chair pressure–Leave house, Refrigerator–Get cold drink, Office chair*
 296 *pressure 3, Bowl cupboard–Prepare dinner, Chair pressure–Prepare dinner*) for providing further analysis on
 297 how the equivalence between real and synthetic datasets may vary depending on the related ADL and
 298 sensor type.

299 Door sensors: *Bedroom door* (ADL: Go to bed)

300 Fig. 3 and Table 2 describe the results obtained from the contrast between synthetic and real SEPA
 301 for *Bedroom door* (ADL: Go to bed). In this case, the two-sided CI for the mean difference between
 302 the real and synthetic SEPA does not contain zero (Fig. 3), suggesting that, in case of "Bedroom door
 303 sensor" (ADL: Go to bed), the SEPA generated by the IE Sim simulator is significantly different from
 304 those emanated from the real environment with a confidence level of 95%. This result is consistent with
 305 the small p -value (0,005) derived from the paired t-test which does not provide good evidence for the
 306 equivalence statement. Specifically, the real SEPA (mean = 4,125 events) was found to be meaningfully
 307 higher than the number of events reported by IE Sim (mean = 1,750 events).

<i>Variable</i>	<i>Mean</i>	<i>Standard dev.</i>	<i>S.E. of the mean</i>
SEPA_Bedroom door_Simulation	1,750	1,165	0,412
SEPA_Bedroom door_Real world	4,125	1,126	0,398
Difference	-2,375	1,685	0,596

Table 2. Comparison test between real and simulated SEPA in Bedroom door (ADL: Go to bed).

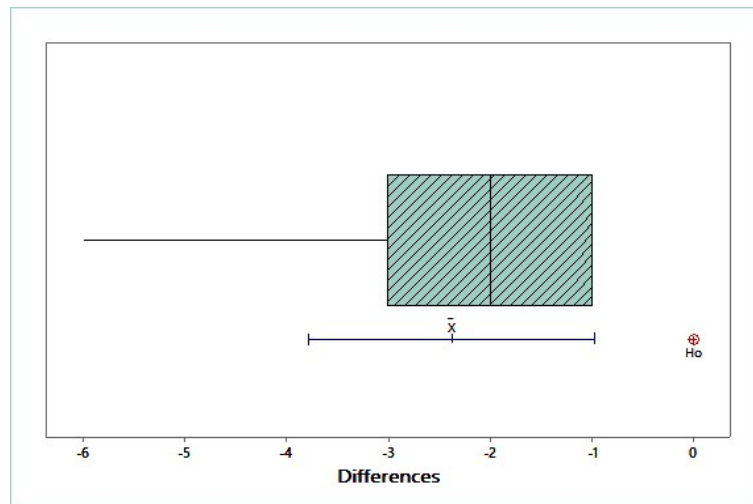


Figure 3. Contrast between real and simulated SEPA – Bedroom door (ADL: Go to bed)

308 The aforementioned analysis was extended to all door sensors so that further insights can be
 309 obtained regarding the equivalence between synthetic data and real observations (refer to Table 3).
 310 In particular, we found that in 83.33% door sensors, the equivalence statement was rejected (p -value
 311 < 0.05). It can be hence inferred that the SEPA produced by the IE Sim simulator and real-world
 312 are considerably dissimilar (CI = 95%). It is hypothesized that the difference between the real and
 313 simulated sensor activation is due to differences in how individuals interact with objects in the real
 314 world versus in the simulation. For example, when entering a room the individual may move through
 315 the doorway and then instinctively push the door closed or partially closed. This would trigger an
 316 activation of the contact sensor. Conversely in the simulation, the user is abstracted from what is
 317 happening in the environment and has to make a concerted effort to interact with each object/ sensor.
 318 In this case, they may sometimes forget to do so. Meaning the door may open but may not be closed.

<i>Sensor–Activity</i>	<i>Two-sided CI for the mean difference between real and simulated SEPA (95%)</i>	<i>t-statistic</i>	<i>p-value</i>	<i>Finding</i>
Bedroom door–Go to bed	[−3,784, −0,966]	−3,99	0,005	SD
Bathroom door–Use bathroom	[−2,066 − 0,184]	−2,83	0,026	SD
Bowl cupboard–Prepare breakfast	[0, 0]	0	1,0	SE
Refrigerator–Prepare breakfast	[−2,682, −0,318]	−3,00	0,020	SD
Refrigerator–Get cold drink	[−1,721, 0,864]	−0,81	0,0448	SD
Bowl cupboard–Prepare dinner	[1,159, 2,341]	−7,00	0,000	SD

Table 3. Contrast between simulated and real door SEPA.

319 Pressure sensors: *Chair pressure* (ADL: Prepare breakfast)

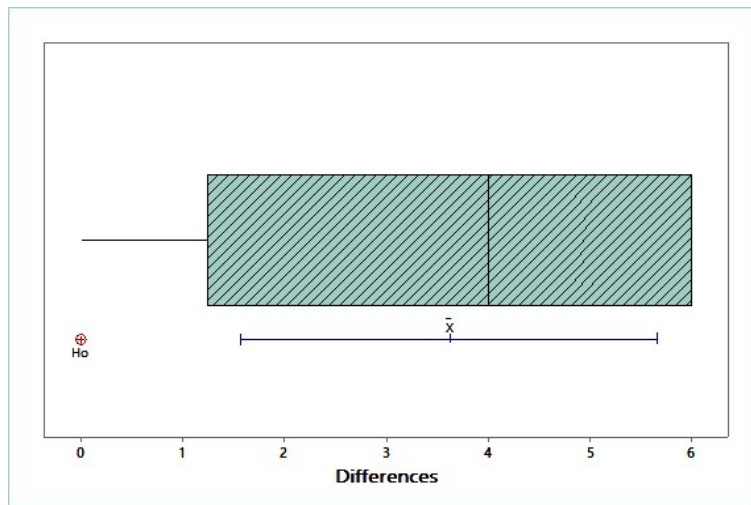


Figure 4. Differences between real and simulated SEPA - Chair pressure (ADL: Prepare breakfast)

320 Fig. 4 and Table 4 present the results of the paired t-test supporting the contrast between the
 321 SEPA (Chair pressure) from IE Sim and the real environment when preparing breakfast. Considering
 322 that the CI for the mean difference between the compared variables does not include zero, there is
 323 then not enough support for the equivalence statement (Fig. 4). This finding is confirmed by the small
 324 p-value (0,004) associated with the null hypothesis, which further suggests (CL = 95%), in this case, no
 325 statistical similarity between real data and those produced using the simulator. In particular, the real
 326 number of SEPA (mean = 1,250 events) was found to be meaningfully lower compared to the SEPA
 327 obtained from the IE simulator (mean = 4,875 events).

<i>Variable</i>	<i>Mean</i>	<i>Standard dev.</i>	<i>Standard error of the mean</i>
SEPA_Chair pressure_Synthetic	4,875	1,885	0,666
SEPA_Chair pressure_Real world	1,250	1,282	0,453
Difference	3,625	2,446	0,865

Table 4. Contrast between real and simulated SEPA in Chair pressure (ADL: Prepare breakfast).

328 The outcomes emanating from the comparative analysis are detailed in Table 5. Supported on
 329 statistical evidence, it was inferred that in 60% of the pressure sensors, the equivalence hypothesis was
 330 rejected. Hence, it can be deduced that the SEPA derived from IE Sim and real-world are different (CI
 331 = 95%) in most sensors.

<i>Sensor–Activity</i>	<i>Two-sided CI for the difference between real and simulated SEPA</i>	<i>t-value</i>	<i>p-value</i>	<i>Conclusion</i>
Bed pressure–Go to bed	[2, 46, 34, 29] (90%)	2, 19	0,065	SD
Chair pressure–Prepare breakfast	[1, 580, 5, 670] (95%)	4, 19	0,004	SD
Chair pressure–Leave house	[–0, 62, 5, 37] (95%)	1, 87	0,103	SE
Office chair pressure 3–Be in the office	[–2, 45, 5, 31] (95%)	0, 90	0,403	SE
Chair pressure–Prepare dinner	[2, 741, 5, 009] (95%)	8, 08	0,000	SD

Table 5. Comparison between simulated and real pressure sensor events per activity.

332 5.2. Predicting real SEPA using simulated data: the application of Poisson regression

333 Given that the equivalence hypothesis was rejected in most door and pressure sensors (sub-section
334 5.1), the next step was to define *How the synthetic data could be modified to better approximate the real SEPA*.
335 Two types of Poisson regression-based models were proposed to deal with this challenge: *sensor-based*
336 and *dummy-variable based*. The following sub-sections will illustrate the results obtained from each of
337 these models including validation (overdispersion, normality, and independence of residuals) and
338 assessment of predictive ability $R\text{-Sq}(adj)$. It is noteworthy that no regression model was defined for
339 Sensor 7: *Bowl cupboard (Prepare dinner)* due to lack of sufficient data.

340 5.2.1. Sensor-based Poisson regression model

341 **Poisson regression models were defined for the above-described sensors by utilizing Minitab 17[®]**
342 **statistical package**. The resulting equations have been validated (see subsections 3.1, 3.2, and 3.3) for
343 ensuring their applicability in practical scenarios. The use of sensor-based models is proposed given
344 the diversity of ADLs considered in this study. As mentioned above, models with high predictive
345 ability can be used for complementing real datasets and then training algorithms capable of recognizing
346 ADLs accurately.

347 Sensor 1: *Bedroom door* (ADL: Go to bed)

348 In this case, the model (9) was **concluded to be statistically significant** ($p\text{-value} = 0.000$) at an
349 alpha level of 0.05. This means that at least one predictor coefficient is different from 0 as noted
350 in Eq. 9. Furthermore, X_1 (simulated events per activity) and interactions including X_2 (simulated
351 activity duration), and X_3 (simulated events per sensor) were found to be significant. Thus, a model
352 considering these terms may be suitable for predicting Y (real events per sensor).

353 In this case, the predictive ability of the model was concluded to be satisfactory ($R_{adj}^2 = 90.21\%$). Such
354 a model was also found to have the lowest Akaike Information Criterion -AIC (35.57) and is then
355 concluded to strike a superior balance between data fit and its ability to tackle overfitting.

356 On the other hand, an Anderson-Darling test was undertaken for assessing the normality of residuals
357 (Fig.5). Considering that $p\text{-value} = 0,338$ and $AD = 0,366$, it can be assumed that the residuals do not
358 deviate significantly from the normal probability distribution. Also, the independence assumption was
359 validated through an auto-correlation test whose metrics ($Max|T| = 0.44$) evidenced no dependence
360 among residuals. Lately, the deviance ($p\text{-value} = 0.829$) and Pearson ($p\text{-value} = 0.846$) coefficients
361 were used to verify the equidispersion phenomenon. Given that both $p\text{-values}$ are higher than the
362 significance level (0.05), this condition is then discarded and the proposed **logarithmic** equation (Eq. 9)
363 can be suggested for predicting the response variable Y .

$$\ln Y = 0.2778X_1 - 0.00678 X_1 * X_2 + 0.000033 X_1 * X_2^2 + 0.000122 X_2^2 * X_3 \quad (9)$$

364 Sensor 2: *Bed pressure* (ADL: Go to bed)

365 Similar to Sensor 1 *Bedroom door* (Go to bed), the predictive model (10) was concluded to be
366 statistically significant ($p\text{-value} = 0$) at 0.05. In this case, X_1 (simulated events per activity), X_2

367 (simulated activity duration), and interactions including X_1 , X_2 , and X_3 (simulated events per sensor)
 368 were identified to explain the response variable. Thus, a Poisson-regression-based model incorporating
 369 these predictors may be appropriate for obtaining new Y (real events per sensor) observations. Indeed,
 370 the predictive ability $R\text{-Sq}(adj)$ was calculated to be 93,24% which ensures reasonable estimations of Y .
 371 AIC index (40.73) also validates this conclusion while confirming the good data fit provided by the
 372 model.

373 On a different tack, the Eq.10 was concluded to satisfy the Poisson regression assumptions. First, the
 374 normality of residuals was **verified** through the Anderson-Darling test statistic ($AD = 0.237$; $p\text{-value} =$
 375 0.688) and Quantile-Quantile plots (Fig.5). Besides, the auto-correlation was estimated to be $\rho(h) = 0$
 376 ($Max|T| = 0.61$) for all $h \neq 0$; thereby supporting the independence of residuals. Ultimately, the
 377 Deviance ($p\text{-value} = 0,721$) and Pearson ($p\text{-value} = 0,716$) statistical tests sustained the equidispersion
 378 assumption. Based on these results, the logarithmic equation is concluded to be valid for predicting
 379 the real number of sensor (Bed pressure) events when *Going to bed*.

$$\ln Y = -1.512X_1 + 0.859 X_2 + 0.1085 X_1^2 - 0.0272 X_2^2 + 0.001705 X_1 * X_2^2 + 0.00355 X_1^2 * X_2 \quad (10)$$

380 Sensor 3: *Bathroom door (Use bathroom)*

381 The Poisson regression modelling (11) was also found to be suitable ($p\text{-value} = 0$) for
 382 approximating the real number of events registered by *Bathroom door* sensor when participants used the
 383 bathroom. Indeed, the good data fit was evidenced through the correlation coefficient ($R_{adj}^2 = 95.09\%$)
 384 and AIC (30.59).

385 The quality of the model described in 11 was verified by assessing the assumptions explained in Section
 386 3. First, the homogeneous property of the Poisson equation was confirmed through the Deviance
 387 ($p\text{-value} = 0.987$) and Pearson ($p\text{-value} = 0.988$) tests which were concluded to accept the null
 388 hypothesis. The normality distribution of model residuals was evaluated using the Anderson-Darling
 389 test and *QQ – plots* (Fig. 5). In this case, the resulting $p\text{-value}$ ($p\text{-value} = 0,756$) and *AD* coefficient
 390 (0.218) provide enough support for the normality assumption. Ultimately, the covariates were
 391 concluded to be uncorrelated within each sample ($Max|T| = 1.0$) which confirms the independence
 392 property of the model. Being aware of the above-mentioned findings, the Eq.11 is concluded to be
 393 valid for predicting the response Y . Of particular interest is the inclusion of X_1 (simulated events per
 394 activity) as the only dependent variable capturing Y variations.

$$\ln Y = 0.2869 X_1 + 0.01313 X_1^2 \quad (11)$$

395 Sensor 4: *Refrigerator (Prepare breakfast)*

396 The use of a Poisson regression model (12) was also appropriate for estimating the real SEPA
 397 when participants opening the refrigerator during breakfast preparation ($p\text{-value} = 0,000$). Such a
 398 finding was checked upon estimating ($R_{adj}^2 = 92.92\%$) and AIC (29.98) which reflect a high performance
 399 concerning prediction ability and data fit respectively.

$$\ln Y = 0.001176 X_2^2 - 0.000013 X_2^3 \quad (12)$$

400 To justify the use of this model in practical scenarios, Poisson regression assumptions were
 401 checked. Initially, the residuals were verified for deviation from normality (Fig.5). In this case, the
 402 results ($p\text{-value} = 0.566$; $AD = 0.271$) are in favour of the normality hypothesis. On the other hand, the
 403 auto-correlation test was performed to detect potential correlations with Y and residuals from previous
 404 models. Given that $Max|T| = 0.82$, dependence among residuals is discarded at a significance
 405 level of 0.05. The assessment of the Poisson model also included evaluating the equidispersion
 406 property. The Deviance's ($p\text{-value} = 0.965$) and Pearson's ($p\text{-value} = 0.965$) goodness-of-fit statistics
 407 concluded against the alternative; therefore, *overdispersion* phenomenon cannot be sustained.

408 Considering the aforementioned results, the model (Eq. 12) is assumed to be appropriate for
 409 predicting the response variable Y . Similar to the previous model, only one variable (X_2 : Simulated
 410 activity duration) was identified as a good predictor of real SEPA in Refrigerator (Prepare breakfast) sensor.
 411

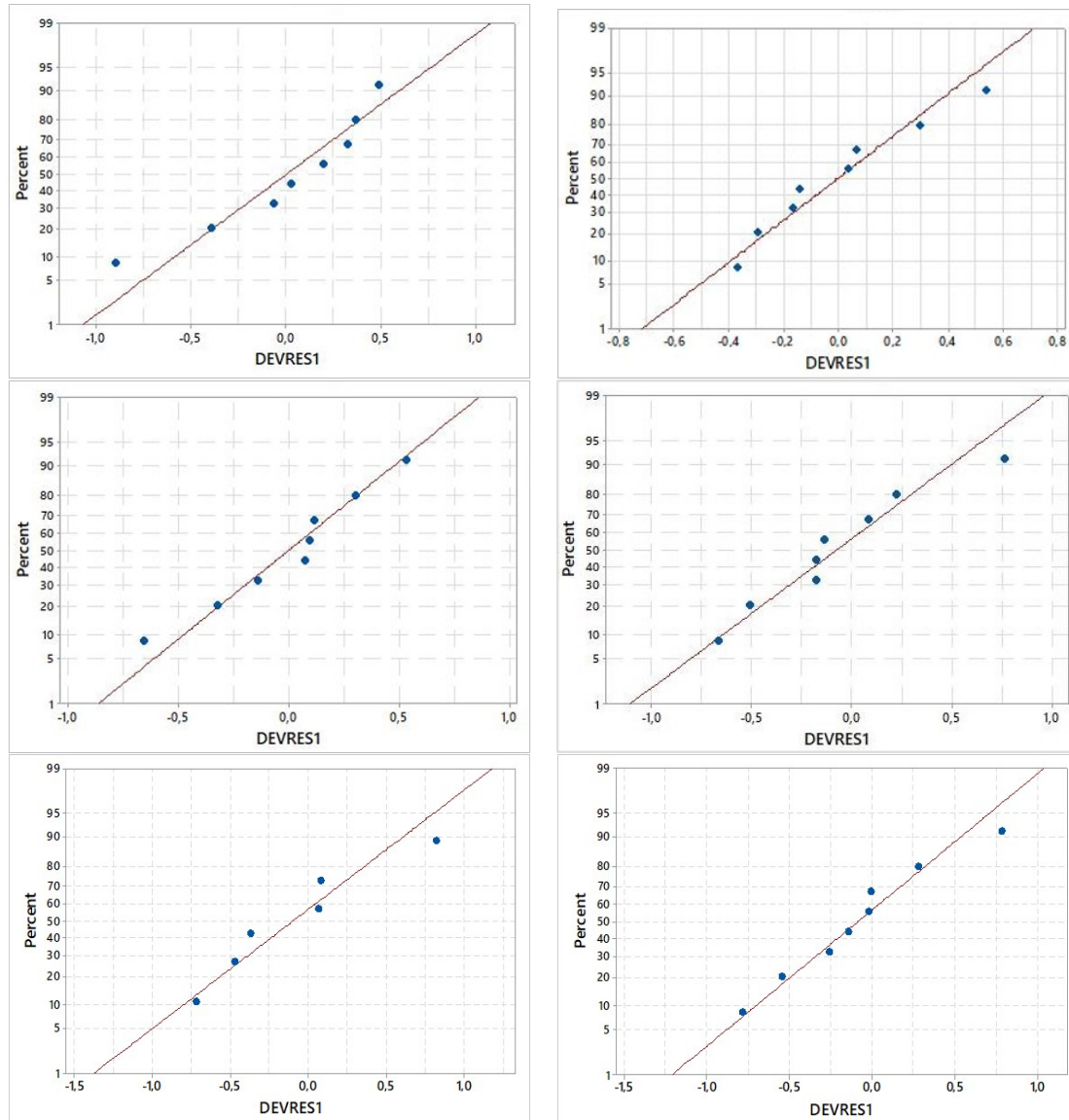


Figure 5. Normality plots of residuals of real SEPA. The sensors are in the first row from the left: Bedroom door (Go to bed), Bed pressure (Go to bed), in the second row: Bathroom door (Use the bathroom), Refrigerator (Prepare breakfast). in third row: Chair pressure (Prepare breakfast), Chair pressure (Prepare dinner)

412 Sensor 5: Chair pressure (Prepare breakfast)

413 In Chair pressure (Prepare breakfast), two interaction terms (including X_1 and X_3) were found
 414 to be significant at 0.05 and 0.1 respectively: $X_1 * X_3$ (p -value = 0.049) and $X_1 * X_3^2$ (p -value = 0.072).
 415 Nevertheless, the best predictive model incorporating these variables (13) (Normality: AD = 0.281
 416 p -value = 0.508 (Fig. 5); Independence: $Max|T| = 1.45$; Equidispersion: Deviance [p -value = 0.816 and
 417 Pearson [p -value = 0.816) was not considered acceptable for estimating the real SEPA in this sensor
 418 ($R_{adj}^2 = 56.36\%$). It is then recommended to include other variables explaining the variability of sensor

419 events when sitting on the chair (ADL: Prepare breakfast). Thereby, the predictive ability of the model
420 can be improved for better training activity recognition algorithms focused on this ADL.

$$\ln Y = 0.0563 X_1 * X_3 - 0.00817 X_1 * X_3^2 \quad (13)$$

421 Sensor 6: *Chair pressure (Prepare dinner)*

422 A p -value = 0.016 confirms that Poisson regression modelling (Eq. 14) is suitable for representing
423 *chair pressure* sensor events upon preparing dinner. In Eq. 14, X_3 (simulated events per sensor) and a
424 quadratic combination between X_1 (SEPA) and X_2 (simulated activity duration) were found to explain
425 part of the response (Y) variability. ($R_{adj}^2 = 73.11\%$) and AIC (11.69) indicate an acceptable ability for
426 predicting the real observations (Y) based on synthetic data (X_1 , X_2 , and X_3).

$$\ln Y = 0.297 X_3 - 0.000889 X_1^2 * X_2 \quad (14)$$

427 The suitability of the model presented in Eq.14 was validated through the normality,
428 independence, and equidispersion assumptions (see Section 3). On one hand, the Anderson-Darling
429 test revealed that residuals follow a normal distribution ($AD = 0.191$; p -value = 0.846). On the other
430 hand, the auto-correlation analysis revealed no interdependence among residuals ($Max|T| = 1.14$).
431 Ultimately, p -values of Deviance (0.946) and Pearson (0.960) tests evidence no overdispersion within
432 the Poisson distribution ($\kappa < 0$). Thus, Eq. 14 provides a reasonable approximation of real sensor
433 events when sitting down at a kitchen chair.

434 Table 6 summarizes the results (prediction performance and validation) of the sensor-based
435 regression models. In this study, most of the models (Eqs. 9, 10, 11, and 12) were found to provide
436 excellent predictions of real SEPA while Eq. 14 and 13 were concluded to offer acceptable and
437 non-satisfactory transformations of synthetic data respectively.

Sensor	Bedroom door	Bed pressure	Bathroom door	Refrigerator	Chair pressure	Chair pressure
ADL	Go to bed	Go to bed	Use bathroom	Prepare breakfast	Prepare breakfast	Prepare dinner
R^2 (adj.)	0.9021	0.9324	0.9509	0.9292	0.5636	0.7311
AIC	35.57	40.73	30.59	29.98	19.44	11.69
Assessment of Poisson regression model						
Auto-correlation T-statistic	0.44	0.61	1.0	0.82	1.45	1.14
Normally distributed residuals p-value	0.338	0.688	0.987	0.566	0.508	0.846
Equidispersion Deviance p-value	0.829	0.721	0.987	0.965	0.816	0.946
Equidispersion Pearson p-value	0.846	0.716	0.988	0.965	0.816	0.960

Table 6. Summary of determination coefficient values and model assessment results for the different sensors.

438 5.2.2. Poisson regression incorporating dummy variables

439 Dummy binary variables were also incorporated into the Poisson regression modelling as an
440 alternative for predicting the real SEPA of any sensor. In this case, these parameters C_i denote the
441 presence or absence of a particular sensor i ($i = 1, 2, \dots, 6$) which is defined by two specific codes: 1 and
442 0 correspondingly. The use of these artificial variables then facilitates the application of a standard
443 predictive model that can be adapted to different types of sensors [39]. The dummy variables to be
444 included in the standard Poisson regression model were predefined as follows:

445 C_1 (Bedroom door - Go to bed): $C_1 = 1$ if the sensor is *Bedroom door - Go to bed*, 0 otherwise.

446

447 C_2 (Bed pressure - Go to bed): $D_2 = 1$ if the sensor is *Bed pressure - Go to bed*, 0 otherwise.

448

449 C_3 (Bathroom door - Use bathroom): $D_3 = 1$ if the sensor is *Bathroom door - Use bathroom*, 0
450 otherwise.

451

452 C_4 (Refrigerator - Prepare breakfast): $C_4 = 1$ if the sensor is *Refrigerator - Prepare breakfast*, 0
453 otherwise.

454

455 C_5 (Chair pressure - Prepare breakfast): $C_5 = 1$ if the sensor is *Chair pressure - Prepare breakfast*, 0
456 otherwise.

457

458 C_6 (Chair pressure - Prepare dinner): $C_6 = 1$ if the sensor is *Chair pressure - Prepare dinner*, 0
459 otherwise.

460

461 In this case, X_1 (SEPA), X_2 (simulated activity duration), X_3 (simulated events per sensor), C_2 and
462 C_5 were included into the model either in a single or combined form. Table 7 presents the variables
463 with significant influence on real SEPA Y . Eq. 15, 16, and 17 (based on three combinations of D_2 and
464 D_5) consolidate these terms for predicting the real SEPA of the sensors.

Predictor	DF	Contribution	Adj MS	F-value	P-value
X_1	1	29.81	29.812	128.25	0.000
X_2	1	3.29	3.296	14.18	0.001
D_2	1	6.13	6.132	26.38	0.000
X_2^2	1	17.71	17.716	76.22	0.000
$X_2 * D_2$	1	4.40	4.405	18.95	0.000
$X_3 * D_5$	1	1.43	1.428	6.15	0.018
X_3^3	1	2.89	2.892	12.44	0.001
Error	35	8.13	0.232		
Total	42	151			

Table 7. ANOVA results for the dummy-variable based model

465 1. $D_2 = 0, D_5 = 0$

466

$$\sqrt{Y} = 0.0512 X_1 + 0.04758 X_2 - 0.000423 X_2^2 + 0.000019 X_1 * X_3^2 \quad (15)$$

467

468

469 2. $D_2 = 0, D_5 = 1$

470

$$\sqrt{Y} = 0.0512 X_1 + 0.04758 X_2 - 0.000423 X_2^2 + 0.000019 X_1 * X_3^2 - 0.1086 X_3 \quad (16)$$

471

472

473 3. $D_2 = 1, D_5 = 0$

474

$$\sqrt{Y} = 2.493 + 0.0512 X_1 - 0.0068 X_2 - 0.000423 X_2^2 + 0.000019 X_1 * X_3^2 \quad (17)$$

475

476

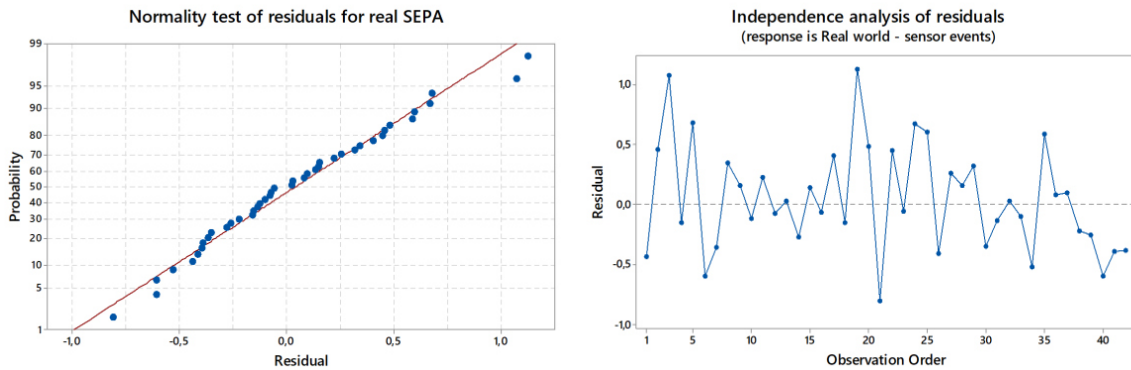


Figure 6. QQ-plot and auto-correlation of real SEPA: Poisson regression model with dummy variables

477

478

479

480

481

482

Table 8 enlists the performance metrics of the above Poisson regression models. Both R^2 (0.9461) and R^2_{adj} (0.9353) values evidence excellent data fit. In a similar vein, the small difference (0.0108) between these coefficients reveals no overfitting problems. On a different tack, R^2_{pred} (0.9272) denotes high predictive ability and new observations can be therefore derived for effectively training activity recognition algorithms. This is confirmed by the error standard deviation and PRESS whose values (0.4821 and 10.9856 respectively) are close to 0.

S	R^2	$R^2 (adj.)$	$R^2 (pred)$	$PRESS$
0.4821	0.9461	0.9353	0.9272	10.9856

Table 8. Performance indicators of the Poisson regression model with dummy variables.

483

484

485

486

487

488

489

490

491

Following this, the normality, equidispersion, and independence properties were assessed for validating the proposed Poisson regression models. Initially, the Anderson-Darling test (see Fig.6) was undertaken for defining whether the residuals follow a normal distribution. The results confirmed the normality hypothesis ($AD = 0.272$; p -value = 0.654) with a mean equals to zero. On the other hand, the randomness test revealed no significant auto-correlations among residuals ($Max|T| = 0.96$). Indeed, Fig.6 does not evidence the presence of runs nor other non-random patterns. Ultimately, both Pearson (p -value > 0.15) and Deviance (p -value > 0.15) were found to confirm the equidispersion property. The aforementioned outcomes confirm the appropriateness of the dummy-variable-based model for their use in the wild.

492

493

494

495

496

497

Upon analyzing the results derived from Poisson regression models, we propose the application of sensor-based models in Go to bed (Bed pressure), Bathroom door (Use bathroom), and Refrigerator (Prepare breakfast). In contrast, the dummy-variable-based model is suggested for predicting the real SEPA derived from the rest of the sensors. The rationale behind this decision is the superior performance provided by this model ($R^2_{pred} = 92.72\%$) compared to those resulting from sensor-based models (90.21%, 56.36%, and 73.11%).

498

6. Conclusions

499

500

501

502

503

504

This paper presented the use of Poisson regression modelling for transforming simulated smart home data to provide an improved approximation of real SEPA. In doing this, synthetic and real data were compared to verify the equivalence hypothesis. This analysis indicated that sensor events per activity produced by the IESim simulator and real-world data do not tend to be statistically equivalent. These results indicate that whilst interactive simulators provide opportunities to facilitate the collection of data in the absence of a real environment, simulated data may not be truly reflective of that collected

505 in the real world.

506 Results indicated that real SEPA can be better approximated ($R^2_{\text{pred}} = 92.72\%$) if synthetic data is
 507 **post-processed** through Poisson regression incorporating dummy variables. Such model is particularly
 508 suggested for predicting the real SEPA derived from three of the sensors cited in this study (Bedroom
 509 door - ADL: Go to bed, Chair pressure - ADL: Prepare breakfast, and Chair pressure - ADL: Prepare
 510 dinner). On a different tack, Eq. 10 ($R^2_{\text{pred}} = 93.24\%$), Eq. 11 ($R^2_{\text{pred}} = 95.09\%$), and Eq. 12 ($R^2_{\text{pred}} =$
 511 92.92%) are recommended for training algorithms recognizing three ADLs: *Go to bed*, *Use bathroom*,
 512 and *Prepare breakfast*. Further, the real SEPA from sensors *Bedroom door*, *Bed pressure*, *Bathroom door*,
 513 *Refrigerator*, *Chair pressure-Prepare breakfast*, and *Chair pressure-Prepare dinner* are well captured by a
 514 combination of Poisson modelling with quadratic (see 9, 10, 11, 14, 15, 16, and 17) and cubic (see 12, 13)
 515 covariates.

516 It is important to note **the** limitations of this study. In particular, the assessment was carried out
 517 using simulated data from one simulator, IESIm. Therefore, it is not possible to tell whether simulated
 518 data **produced by other simulators would generate the same results**. It is also not possible to say
 519 whether similar techniques would work with model-based approaches to data simulation. Nonetheless,
 520 results from this research highlight the importance of considering the quality of simulated data when
 521 modelling solutions for human activity recognition. Future work will investigate the applicability
 522 of these findings to data generated by other simulation techniques including both interactive and
 523 **model-based** approaches.

524 **One appealing idea is the ability to use activity data from one person and transform these so that**
 525 **they fit according to the profile of another person. A means to this end could be achievements from**
 526 **the theory of transfer learning [40]. The relationships making this kind of transformation possible has**
 527 **not been covered in this study but remains as an urgent question for future research.**

528
 529 **Author Contributions:** “Conceptualization, M.O. and E.J.; methodology, M.O. and E.J.; software, M.O. and J.S.;
 530 validation, M.O.; formal analysis, M.O. and J.S.; investigation, J.S., M.O., P.P., C.N., I.C. and E.J.; resources, C.N., I.C.
 531 and M.O.; data curation, M.O.; writing—original draft preparation, M.O., I.C., P.P., C.N. and E.J.; writing—review
 532 and editing, M.O., I.C., P.P., C.N., E.J. and J.S.; visualization, M.O. and E.J.; supervision, C.N., I.C. and M.O.;
 533 project administration, C.N., M.O. and I.C.; funding acquisition, C.N. and I.C.”.

534 **Funding:** This research has received funding under the REMIND project Marie Skłodowska-Curie EU Framework
 535 for Research and Innovation Horizon 2020, under Grant Agreement No. 734355.

536 **Acknowledgments:** The authors would like to thank Giselle Paola Polifroni Avendaño for her valuable support
 537 during this research. **Also many thanks to Jens Lundström for his helpful comments and previous work in the**
 538 **field.**

539 **Conflicts of Interest:** The authors declare no conflict of interest.

540 Abbreviations

541 The following abbreviations are used in this manuscript:

542	AD	Anderson-Darling
	ADL	Activities of Daily Living
	CI	Confidence interval
	GP	Generalized Poisson
	HINT	Halmstad Intelligent Home
	IE Sim	Intelligent Environmental Simulation
543	IoT	Internet of Things
	PIR	Passive Infrared
	QQ	Quantile-Quantile
	RFID	Radio Frequency Identification
	SD	Statistically different
	SE	Statistically equivalent
	SEPA	Sensor events per activity

544 **References**

- 545 1. Ortiz, M.A.; López-Meza, P. Using computer simulation to improve patient flow at an outpatient internal
546 medicine department. *International Conference on Ubiquitous Computing and Ambient Intelligence*.
547 Springer, 2016, pp. 294–299.
- 548 2. Barrios, M.A.O.; Caballero, J.E.; Sánchez, F.S. A methodology for the creation of integrated service networks
549 in outpatient internal medicine. *Ambient intelligence for health*. Springer, 2015, pp. 247–257.
- 550 3. Cheng, L.; Nugent, C.D. *Human Activity Recognition and Behaviour Analysis*, 1 ed.; Springer Nature
551 Switzerland, 2019; chapter Sensor-Based Activity Recognition Review.
- 552 4. Ortiz-Barrios, M.A.; Herrera-Fontalvo, Z.; Rúa-Muñoz, J.; Ojeda-Gutiérrez, S.; De Felice, F.; Petrillo, A.
553 An integrated approach to evaluate the risk of adverse events in hospital sector: from theory to practice.
554 *Management Decision* **2018**, *56*, 2187–2224.
- 555 5. Rafferty, J.; Nugent, C.D.; Liu, J.; Chen, L. From activity recognition to intention recognition for assisted
556 living within smart homes. *IEEE Transactions on Human-Machine Systems* **2017**, *47*, 368–379.
- 557 6. Nugent, C.; Synnott, J.; Gabrielli, C.; Zhang, S.; Espinilla, M.; Calzada, A.; Lundstrom, J.; Cleland, I.;
558 Synnes, K.; Hallberg, J.; others. Improving the quality of user generated data sets for activity recognition.
559 In *Ubiquitous Computing and Ambient Intelligence*; Springer, 2016; pp. 104–110.
- 560 7. Helal, S.; Kim, E.; Hossain, S. Scalable approaches to activity recognition research. *Proceedings of the 8th*
561 *International Conference Pervasive Workshop*, 2010, pp. 450–453.
- 562 8. Barrios, M.O.; Jiménez, H.F.; Isaza, S.N. Comparative analysis between ANP and ANP-DEMATEL for six
563 sigma project selection process in a healthcare provider. *International Workshop on Ambient Assisted*
564 *Living*. Springer, 2014, pp. 413–416.
- 565 9. Barrios, M.O.; Jiménez, H.F. Reduction of average lead time in outpatient service of obstetrics through six
566 sigma methodology. *Ambient Intelligence for Health*. Springer, 2015, pp. 293–302.
- 567 10. Tapia, E.M.; Intille, S.S.; Larson, K. Activity recognition in the home using simple and ubiquitous sensors.
568 *International conference on pervasive computing*. Springer, 2004, pp. 158–175.
- 569 11. Cook, D.; Schmitter-Edgecombe, M.; Crandall, A.; Sanders, C.; Thomas, B. Collecting and disseminating
570 smart home sensor data in the CASAS project. *Proceedings of the CHI workshop on developing shared*
571 *home behavior datasets to advance HCI and ubiquitous computing research*, 2009, pp. 1–7.
- 572 12. Van Kasteren, T.; Noulas, A.; Englebienne, G.; Kröse, B. Accurate activity recognition in a home setting.
573 *Proceedings of the 10th international conference on Ubiquitous computing*. ACM, 2008, pp. 1–9.
- 574 13. Alshammari, N.; Alshammari, T.; Sedky, M.; Champion, J.; Bauer, C. Openshs: Open smart home simulator.
575 *Sensors* **2017**, *17*, 1003.
- 576 14. De-La-Hoz-Franco, E.; Ariza-Colpas, P.; Quero, J.M.; Espinilla, M. Sensor-based datasets for human activity
577 recognition—A systematic review of literature. *IEEE Access* **2018**, *6*, 59192–59210.
- 578 15. Rafferty, J.; Synnott, J.; Nugent, C.D.; Ennis, A.; Catherwood, P.A.; McChesney, I.; Cleland, I.; McClean, S.
579 A Scalable, Research Oriented, Generic, Sensor Data Platform. *IEEE Access* **2018**, *6*, 45473–45484.
- 580 16. Synnott, J.; Nugent, C.; Jeffers, P. Simulation of smart home activity datasets. *Sensors* **2015**, *15*, 14162–14179.
- 581 17. Lundström, J.; Synnott, J.; Järpe, E.; Nugent, C.D. Smart home simulation using avatar control and
582 probabilistic sampling. *2015 IEEE International Conference On Pervasive Computing And Communication*
583 *Workshops (Percom Workshops)*. IEEE, 2015, pp. 336–341.
- 584 18. Ortiz-Barrios, M.; Lundström, J.; Synnott, J.; Järpe, E.; Sant’Anna, A. Complementing real datasets with
585 simulated data: a regression-based approach. *Multimedia Tools and Applications*, pp. 1–24.
- 586 19. Schreiber, T.; Schmitz, A. Surrogate time series. *Physica D: Nonlinear Phenomena* **2000**, *142*, 346–382.
- 587 20. Maiwald, T.; Mammen, E.; Nandi, S.; Timmer, J. Surrogate data—A qualitative and quantitative analysis.
588 In *Mathematical Methods in Signal Processing and Digital Image Analysis*; Springer, 2008; pp. 41–74.
- 589 21. Salazar, A.; Safont, G.; Vergara, L. Surrogate techniques for testing fraud detection algorithms in credit card
590 operations. *2014 International Carnahan Conference on Security Technology (ICCST)*. IEEE, 2014, pp. 1–6.
- 591 22. Abroug, F.; Ouanes-Besbes, L.; Elatrous, S.; Brochard, L. The effect of prone positioning in acute respiratory
592 distress syndrome or acute lung injury: a meta-analysis. *Areas of uncertainty and recommendations for*
593 *research*. *Intensive care medicine* **2008**, *34*, 1002.

- 594 23. Synnott, J.; Chen, L.; Nugent, C.D.; Moore, G. The creation of simulated activity datasets using a
595 graphical intelligent environment simulation tool. 2014 36th Annual International Conference of the IEEE
596 Engineering in Medicine and Biology Society. IEEE, 2014, pp. 4143–4146.
- 597 24. Ariani, A.; Redmond, S.J.; Chang, D.; Lovell, N.H. Simulation of a smart home environment. 2013 3rd
598 International Conference on Instrumentation, Communications, Information Technology and Biomedical
599 Engineering (ICICI-BME). IEEE, 2013, pp. 27–32.
- 600 25. Francillette, Y.; Boucher, E.; Bouzouane, A.; Gaboury, S. The Virtual Environment for Rapid Prototyping of
601 the Intelligent Environment. *Sensors* **2017**, *17*, 2562.
- 602 26. Park, B.; Min, H.; Bang, G.; Ko, I. The User Activity Reasoning Model in a Virtual Living Space Simulator.
603 *International Journal of Software Engineering and Its Applications* **2015**, *9*, 53–62.
- 604 27. Lee, J.W.; Cho, S.; Liu, S.; Cho, K.; Helal, S. Persim 3d: Context-driven simulation and modeling of human
605 activities in smart spaces. *IEEE Transactions on Automation Science and Engineering* **2015**, *12*, 1243–1256.
- 606 28. McGlenn, K.; O’Neill, E.; Gibney, A.; O’Sullivan, D.; Lewis, D. SimCon: A Tool to Support Rapid
607 Evaluation of Smart Building Application Design using Context Simulation and Virtual Reality. *J. UCS*
608 **2010**, *16*, 1992–2018.
- 609 29. Renoux, J.; Klugl, F. Simulating daily activities in a smart home for data generation. 2018 Winter Simulation
610 Conference (WSC). IEEE, 2018, pp. 798–809.
- 611 30. Mendez-Vazquez, A.; Helal, A.; Cook, D. Simulating events to generate synthetic data for pervasive spaces.
612 Workshop on Developing Shared Home Behavior Datasets to Advance HCI and Ubiquitous Computing
613 Research. Citeseer, 2009.
- 614 31. Cameron, A. *Regression analysis of count data*; Cambridge University Press, 1998.
- 615 32. Kunkler, M. Modelling negatives in stochastic reserving models. *Insurance: Mathematics and Economics*
616 **2006**, *38*.
- 617 33. Andersson, P.K.; Skovgaard, L.T. *Regression with Linear Predictors*; Springer, 2010.
618 doi:10.1007/978-1-4419-7170-8.
- 619 34. Joe, H.; Zhu, R. Generalized Poisson distribution: the property of mixture of Poisson and comparison with
620 negative binomial distribution. *Biometrical Journal* **2005**, *47*, 219–229.
- 621 35. Consul, P.; Famoye, F. Generalized Poisson regression-model. *Communications in statistics – Theory and*
622 *methods* **1992**, *21*, 89–109.
- 623 36. Marsaglia, G. Evaluating the Anderson-Darling Distribution. *Journal of Statistical Software* **2005**, *9*, 219–229.
- 624 37. Ljung, G.; Box, G. On a Measure of a Lack of Fit in Time Series Models. *Biometrika* **1978**, *65*, 297–303.
625 doi:10.1093/biomet/65.2.297.
- 626 38. Lundström, J.; De Morais, W.O.; Menezes, M.; Gabrielli, C.; Bentes, J.; Sant’Anna, A.; Synnott, J.; Nugent, C.
627 Halmstad intelligent home-capabilities and opportunities. International Conference on IoT Technologies
628 for HealthCare. Springer, 2016, pp. 9–15.
- 629 39. Nisbet, R.; Elder, J.; Miner, G. *Handbook of statistical analysis and data mining applications*; Academic Press,
630 2009.
- 631 40. Torrey, L.; Shavlik, J. *Transfer learning. Handbook of Research on Machine Learning Applications and Trends:*
632 *Algorithms, Methods, and Techniques*; IGI Global, 2009. doi:10.4018/978-1-60566-766-9.