IPOP-CMA-ES and the Influence of Different Deviation Measures for Agent-Based Model Calibration

1st Víctor Vargas-Pérez
University of Granada
18014 Granada, Spain
vavp98@correo.ugr.es

2nd Manuel Chica
Andalusian Research Institute DaSCI
University of Granada
18014 Granada, Spain
manuelchica@ugr.es
School of Electrical Engineering and Computing
The University of Newcastle
Callaghan, NSW 2308, Australia

3rd Óscar Cordón
Andalusian Research Institute DaSCI
University of Granada
18014 Granada, Spain
ocordon@decsai.ugr.es

***Abstract*—Calibration is a crucial task on building valid models before exploiting their results. This process consists of adjusting the model parameters in order to obtain the desired outputs. Automatic calibration can be performed by using an optimization algorithm and a fitness function, which involves a deviation measure to compare the time series coming from the model. In this paper, we apply a memetic IPOP-CMA-ES for the calibration of an agent-based model and we study the effect of different deviation measures in this calibration problem. Classical metrics calculate the mean point-to-point error, but we also propose using an extension of dynamic time warping, which considers trend series evolution. In order to determine if calibrating with an specific metric leads to better solutions, we carry out an exhaustive experimentation by including statistical tests, analysis on the values of the calibrated parameters, and qualitative results. Our results show IPOP-CMA-ES obtains better performance than a genetic algorithm. In addition, MAE, MAPE and Soft-DTW are the metrics which report best results, although we get a similar behavior for all of them.**

***Keywords*—model calibration, agent-based modeling, time series, deviation measures, evolutionary algorithms**

I. Introduction

Agent-Based Modeling (ABM) [1] is a technique that allows to recreate complex systems present in the real world. This recreation is performed through an artificial agent population which is usually connected with a network. These agents are defined by a set of micro-rules and the interaction between them generates the emergent behavior or macro-effects that are found in the reality. ABM can be used in many diverse fields, from ecology [2], sociology [3], epidemiology [4] to marketing [5]. In the latter case, i.e., marketing, ABM allows to learn the dynamics between clients and brands, and test campaigns (what-if scenarios) on a virtual market without the risk of launching them into the real market.

However, building an agent-based model requires a phase of calibration, where a large set of unknown parameters are needed to be adjusted to obtain outputs similar to a given historical data. This calibration can be performed automatically using three elements: an optimization algorithm (such as evolutionary algorithms), historical data, and a fitness function that compares output model with the given historical data. The algorithm will modify the model parameters systematically, with the goal of reduce the fitness function value, which is defined as a deviation measure.

Previous works show the complexity of this calibration process [6]. In [7], authors showed the importance of multimodal optimization algorithms, which return different suboptimal solutions that can be evaluated by an expert. On the other hand, [8] addressed the calibration as a multiobjective problem: agent-based models usually have multiple outputs and the minimization of the error for all of them can be in conflict. In addition, there is a need to validate the model to ensure that its behavior corresponds to reality. As shown in [9], this procedure involves aspects as determining minimum simulations runs, a sensitive analysis of the parameters, and reviewing the spatio-temporal dynamics of the model outputs. Since both historical data and each output of an agent-based model take the form of time series with the value for a measure at each time step of the simulation, the fitness function involves a deviation measure to compare both streams of data.

In this paper, we calibrate a marketing agent-based model with a memetic variant of IPOP-CMA-ES, a sophisticated evolutionary algorithm which has a competitive performance [10]. In addition, this algorithm has shown good results for calibration purposes [11]. We also perform an extensive study of different metrics in the automatic calibration of this agent-based model. Classical metrics measure the point-to-point error (such as MAE or RMSE), but they are not capable of considering the trend evolution of the series. For this reason, we propose the use of the algorithm Soft dynamic time warping (Soft-DTW) [12], which does consider the trends. This algorithm is a modification of dynamic time warping (DTW) [13] which is more suitable to average and cluster time series. As we shall see, the evaluation of a solution requires averaging the error for different stochastic simulations. Therefore, Soft-DTW could be an appropriate choice for the calibration problem and was not used before for this task.

We use an agent-based model which simulates a marketing scenario where different brands invest in a set of touchpoints to improve key performance indicators such as brand awareness. Awareness values of each consumer agent can be altered by word of mouth interactions with their neighbors and touchpoints advertising. We will calibrate some of the parameters that model these mechanisms.

To carry out this study, we will perform an exhaustive experimentation that involves multiple calibration experiments of different instances of this agent-based model. We will compare IPOP-CMA-ES, a stationary genetic algorithm and a basic random search by also comparing four classical metrics and the Soft-DTW. Next, we will perform a *post-hoc* analysis of the calibration results. We will start this analysis by comparing point-to-point metrics through cross-comparison, an approach also used in [6]. Then, we will compare the results for all the metrics (including Soft-DTW) with ranking statistical tests. Finally, we will analyze model's instances outputs and parameters obtained by each metric, showing the boxplots of the parameters and the global outputs of the calibrated models.

The rest of the paper is structured as follows. In Section II, we discuss the use of evolutionary algorithms in ABM calibration and describe the memetic IPOP-CMA-ES. Section III describes all the metrics considered for time series comparison. Then, Section IV depicts the marketing agent-based model to be used for the calibration scenario. Section V shows the whole experimentation with their results and analysis. Finally, we present our conclusions in Section VI

II. Optimization Methods

A. Evolutionary Algorithms for ABM Calibration

Is is necessary to validate agent-based models to guarantee that they properly represent the real system to simulate. A crucial step in model validation is calibration, whose purpose is tuning the unknown parameters of the model to obtain outputs similar to historical data. This process can be perform manually such as a global sensitivity analysis [14]. However, this approach is unfeasible for realistic model, where there are a high number of parameters to calibrate. Another common approach is automatic calibration, where an optimization algorithm is employed [7]. We can find applications of automatic calibration in [15], [16], [5] and [17].

Since the parameters in agent-based models exhibit non-linear interactions, the best option is to use non-linear optimization algorithms such as metaheuristics, which can provide good solutions in a reasonable time. There are multiple examples of metaheuristic application for model calibration, from genetic algorithms [18] to evolution strategies [19] or differential evolution [15].

Evolutionary algorithms are a popular family of metaheuristics [20]. This algorithms work with a population of solutions and, at each iteration, the best solutions are selected as parents to generate new child solutions through a process of crossover and a possible random mutation. Thus, these new solutions will replace the previous population totally or partially, and the cycle will repeat again. A simple example of this algorithms is steady state genetic algorithm (SSGA), focused on elitism: at each iteration, it is selected only a pair of parent solutions, which will generate two child solutions. The new solutions will only replace the worst existing ones if they have better fitness values.

In our case, a solution is a configuration setting for the agent-based model where we evaluate its quality (fitness) through the comparison of its outputs with historical data. Thus, we define the fitness function as follows:

$$f = \frac{1}{|B|} \sum_{b=1}^{|B|} error(h^b, s^b) \tag{1}$$

where B is the set of model outputs, h^b is the objective value (historical data) for output b, s^b is the value obtained for that output b and $error(h^b, s^b)$ is the deviation measure error value of s^b with respect to h^b.

B. IPOP-CMA-ES

Covariance Matrix Adaptation Evolution Strategy (CMA-ES) [21] is an evolutionary algorithm that uses a covariance matrix to create new solutions. At each iteration, λ new individuals are generated independently through a multivariate normal distribution, which is defined as follows:

$$\boldsymbol{x}_k^{(g+1)} \sim \mathcal{N}\left(\langle \boldsymbol{x} \rangle_{\mathrm{w}}^{(g)}, {\sigma^{(g)}}^2 \boldsymbol{C}^{(g)}\right), \quad k = 1, \ldots, \lambda \tag{2}$$

where $\mathcal{N}(\boldsymbol{m}, \boldsymbol{C})$ is a normally distributed random vector with mean m and covariance matrix C.

$\boldsymbol{x}_k^{(g+1)}$ is the k-th individual of generation $g+1$, and $\langle \boldsymbol{x} \rangle_{\mathrm{w}}^{(g)}$ is a weighted mean vector obtained from the μ best individuals of the previous population. Such weighted mean is computed from a weight vector w with μ elements which verifies $\sum_{i=1}^{\mu} w_i = 1$

It is worth noting that the mean $\langle \boldsymbol{x} \rangle_{\mathrm{w}}^{(g)}$ corresponds to the crossover/exploitation mechanism, since its values depend on the best solutions found at each iteration; while mutation/exploration mechanism depends on $\sigma^{(g)}$ (step size) and $\boldsymbol{C}^{(g)}$ (covariance matrix). In order to update this matrix at each iteration, the algorithm uses the fitness function values obtained in previous iterations (evolution path $\boldsymbol{p}_c^{(g)}$) and the last solutions values created.

IPOP-CMA-ES [22] is an extended version of CMA-ES that includes a restart strategy with increasing population. The stopping criterion for each restart is related with the evolution stagnation, considering factors such as the improvement over several generations.

In this work, we use a memetic version of IPOP-CMA-ES due to, as discussed previously, it is a powerful evolutionary algorithm with good results in real-coded optimization competitions [10] and, in particular, its previous use in ABM calibration [11]. We add an additional exploitation mechanism to IPOP-CMA-ES: our variant includes a local search of 50 evaluations over population at each iteration with a probability $p_{LS} = 0.0625$.

III. Deviation Measures for Time Series Comparison

We define the fitness function for ABM calibration in 1. However, since each agent-based model output a_s^b and historical data a_h^b take the form of time series, different deviation measures can be considered as the generic $error$. In this section, we present five deviation measures alternatives, from simple point-to-point metrics (Section III-A) to a more sophisticated metric based on DTW to consider trend evolution (Section III-B).

A. Point-to-point Metrics

We use four classical point-to-point metrics. If *n* is the length of the time series, *s* is the output of the agent-based model and *h* is the historical data, we can define first three metrics as follows:

- Mean Absolute Error (MAE):

$$\mathrm{MAE} = \frac{1}{n}\sum_{j=1}^{n}|h_j - s_j| \tag{3}$$

- Root Mean Squared Error (RMSE):

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(h_j - s_j)^2} \tag{4}$$

- Mean Absolute Percentage Error (MAPE):

$$\mathrm{MAPE} = \frac{1}{n}\sum_{j=1}^{n}\left|\frac{h_j - s_j}{h_j}\right| \tag{5}$$

The fourth point-to-point metric we consider is the coefficient of determination R^2. This is a typical measure in statistics that evaluate how well the model replicates the observed outcomes, based on the proportion of the variance of outcomes explained. It is defined as follows:

$$R^2 = 1 - \frac{\sum_{j=1}^{n}(h_j - s_j)^2}{\sum_{j=1}^{n}\left(h_j - \bar{h}\right)^2} \tag{6}$$

$R^2 \in [0,1]$, where $R^2 = 1$ indicates a perfect fit and $R^2 = 0$ a null fit. However, we need a deviation measure to minimize, therefore we use $1 - R^2$. Thus, ϵ_{R^2} is:

$$\epsilon_{\mathrm{R}^2} = 1 - R^2 = \frac{\sum_{j=1}^{n}(h_j - s_j)^2}{\sum_{j=1}^{n}\left(h_j - \bar{h}\right)^2} \tag{7}$$

This measure is similar to RMSE since the minimization of both metrics only depends on the sum of the squared errors (the denominator is constant in both cases). In addition, although theoretically this measure is defined in the interval $[0,1]$, it is possible to get an error higher than 1 if we consider the previous expression. R^2 evaluate the fit of the model with respect to a horizontal line, therefore if the fit is worse than that, R^2 will be negative and ϵ_{R^2} will be higher than 1.

B. Soft Dynamic Time Warping

In order to consider trend evolution of time series, [9] proposes using DTW [13]. This algorithm measures similarity between two time series, even if these series have different length. Unlike previous metrics, the evaluation is not affected by dilations or displacements on the time dimension. To achieve this, DTW performs the optimal alignment between the series. It means the sequence of pairs of points under certain restrictions with minimum euclidean distance. The time complexity of this algorithm is O(nm), where n and m are the length of each time series. We consider series with the same length, therefore the time complexity is O(n^2), while it is linear in the case of the previous metrics. Algorithm 1 shows the DTW pseudocode.

Algorithm 1 *Dynamic Time Warping*

Input: *Series1, Series2* : arrays with the values of the time series to compare
Output: Distance : Distance bewteen the series

```
n := length(Series1), m := length(Series2)
DTW := array [0..n-1], [0..m-1]
DTW[0, 0] := distance(Series1[0], Series2[0])
for i := 1 to n-1 do
    DTW[0, i] := DTW[0, i-1] + distance(Series1[0], Series2[i])
for i := 1 to m-1 do
    DTW[i, 0] := DTW[i-1, 0] + distance(Series1[i], Series2[0])
for i := 1 to n-1 do
    for j := 1 to m-1 do
        DTW[i, j] := distance(Series1[i], Series2[j]) + min(DTW[i-1, j], DTW[i-1, j-1], DTW[i, j-1])
return DTW[n-1, m-1]
```

As previously mentioned, we use the smooth formulation Soft-DTW [12] because is more suitable to average time series, and we will have to average the error for different time series. This alternative redefines the minimum function as done in 8. This formulation uses a new parameter γ set to $\gamma = 0.001$, as it is the value showing the best results in [12].

$$\min_\gamma\{a_1,\ldots,a_n\} = \begin{cases} \min_{i\le n} a_i, & \gamma = 0 \\ -\gamma\log\sum_{i=1}^{n} e^{-a_i/\gamma}, & \gamma > 0 \end{cases} \tag{8}$$

Fig. 1 illustrates the potential of Soft-DTW with respect to point-to-point metrics. Here, the objective time series is given by the function $cos(x) + 10$ (shown as a blue curve in the figure), and we have two outputs: a constant series centered in 10 (orange line in the figure) and an identical time series to the objective with a displacement $\frac{\pi}{2}$ in time axis (green curve in the figure). It seems reasonable choosing the green output as it is the most similar to the objective. However, as we can see in Table I, all the point-to-point metrics provide a higher error for that series. Soft-DTW is the only metric considering Series 2 as the best solution.

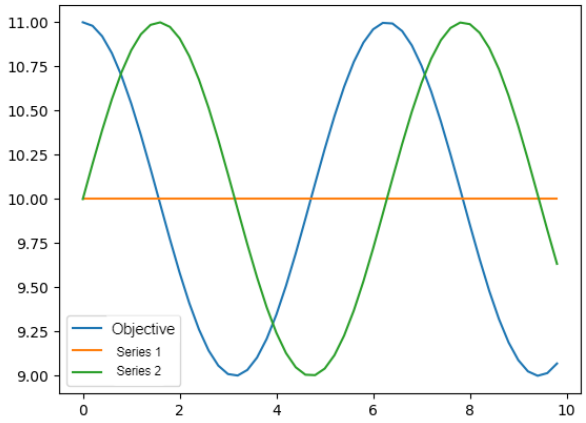

Fig. 1: Two series and an objective function to illustrate the differences of traditional deviation measures with respect to Soft-DTW

TABLE I: Different metrics errors for series in Fig. 1

Series	MAE	MAPE	RMSE	R2	Soft-DTW
Series1	0.656	6.641	0.725	100.245	5.124
Series2	0.894	9.089	0.990	186.940	2.042

IV. The ABM Calibration Problem

A. Agent-Based Model General Description

We use the marketing agent-based model described in [8] and [11]. This model simulates W weeks of a market composed by a set of brands B and a network of N agents. These agents interact with each other, and in turn they are exposed to the media campaigns of a set of touchpoints T. The output involves one key performance indicator for each brand: brand awareness.

All agents have binary variables $a_i^b \in \{0,1\}$, which indicate agent i aware of brand b. The values of these variables are dynamic: at each time step of the simulation, every agent can gain awareness for a brand through the interaction with its neighbors or through an advertisement, and can lose it through a deactivation process if it is not reinforced. This deactivation is modeled with a parameter called awareness decay d.

Each agent i has a talking probability $p_i^b(w) \in [0,1]$ to spread to its neighbors its awareness value for each brand b at time step w. Each neighbor will gain awareness of a brand with a probability defined by the word-of-mouth (WOM) awareness impact parameter $\alpha^{WOM} \in [0,1]$.

Touchpoints act as brand advertising, and they can influence any agent of the model. The maximum number of reached agents depends on the properties of the touchpoint and the amount invested by the brand. Each touchpoint t has a reach parameter $r_t \in [0,1]$ that indicates the maximum number of agents that touchpoint can influence in a step, and an awareness impact parameter α_t that determines the probability of the agent to activate its awareness of the ad brand after an impact. In addition, advertising can create a viral effect in the reached agent, which is modeled by the parameters buzz increment τ_t and buzz decay $d\tau_t$.

B. Calibration Scenario

We calibrate the three touchpoints parameters buzz increment τ_t, buzz decay $d\tau_t$ and awareness impact α_t; and the three social parameters awareness decay d, talking probability $p_i^b(0)$ and WOM awareness impact α^{WOM}. Hence, the number of parameters to calibrate depends on the number of touchpoints. Specifically, the number of parameters is $3|T|+3 = 3(|T|+1)$. All these parameters follow a real codification in range $[0,1]$.

As previously defined, we consider the awareness as our key performance indicator. However, the agent-based model has the same number of outputs as number of brands. Therefore, the fitness function is the mean error for all the brands. Thus, in our calibration problem, each output b of the fitness function 1 corresponds to the awareness of each brand.

In addition, the model output is different at each simulation due to its stochastic components. In order to deal with stochasticity, we perform 10 Monte Carlo simulations at each evaluation. Thus, the final fitness of a solution is the mean of the fitness function value for each of these simulations.

We use six different instances of the agent-based model previously described. This instances were generated artificially from a base instance defined in [11]. This base instance represents a real banking marketing scenario with $|B| = 8$ brands, $|T| = 7$ touchpoints, a population of $N = 1,000$ agents and a simulation time of $W = 52$ weeks.

Each new instance adds new touchpoints - 40, 45 and 50 - through modifications of the original touchpoints. In this way, we increased the number of parameters to calibrate to 144, 159, and 174. Furthermore, in order to obtain achievable historical data, these are generated through simulations with specific parameters. Thus, we have for each problem size two types of historical data: one with multiple peaks (decay) and others with a flat evolution (medium).

V. Experimentation

A. Experimental Setup

We have performed different calibrations on these instances. For all experiments, the calibration with each configuration is repeated 20 times, due to the stochastic elements of search algorithms. On the other hand, as previously discussed, the evaluation of each solution is the mean of the fitness function values upon 10 Monte Carlo simulations. The stopping criteria for all calibrations is reaching 10,000 evaluations.

Since [11] manages a similar problem, we use the same parameters for IPOP-CMA-ES:

- Initial population size $\lambda = 15$
- Numbers of individuals to select $\mu = 6$
- Increasing factor of 2 at each restart
- Learning rates (parameters to update matrix covariance) of $c_\sigma = 0.568$, $c_c = 0.6962$ and $c_{cov} = 0.4897$.
- Initial step size $\sigma^{(0)} = 0.056747$. Reference [11] sets $\sigma^{(0)} = 56.747$ because they use integer codification in

interval [0, 1000]. However, we use real codification in interval [0, 1].

- Dampening for the step size $d_\sigma = 4.2939$.

In addition, we consider a SSGA with a population of 100 individuals, tournament selection with 3 individuals, the blend crossover operator (BLX-α) and a mutation probability of 0.1

B. Comparison of Evolutionary Algorithms using MAPE

First, we compare the results of calibrating with IPOP-CMA-ES, SSGA, and a random search in order to reinforce the choice of IPOP-CMA-ES and to check that this calibration process is not trivial, which would reduce the importance of the metric used. We use MAPE as the main metric for all these algorithms.

Table II shows the summary results for this optimization problem. The quality of the solutions depends on the technique used, and even with IPOP-CMA-ES, the best technique, there is a certain margin for improvement, particularly for the 50TPDEC instance. Thus, the comparative between different metrics gains importance, since the metric is the guide for the optimization algorithm.

C. Comparison of point-to-point metrics and Soft-DTW

We now calibrate all the agent-based model instances with different metrics using IPOP-CMA-ES. Since the four classical metrics have a similar definition, it is interesting to perform a comparative between them to set if some metric dominates the rest. Since interpretation of each measure differs, the direct comparison of the error values is not appropriate. We follow the approach proposed in [6], where the authors study the quality of different metrics in calibration performing a cross-comparison. It means that the solution obtained through the calibration with each metric is evaluated with the rest of them, checking if there are situations where the minimum error obtained for a metric is achieved by calibrating with another metric.

Table III shows that, generally, the best mean and minimum value for each metric is achieved through the calibration with that metric. However, there are some exceptions. For example, calibration with RMSE obtains the minimum MAE value for 40TPMedium instance, while calibration with MAE obtains the minimum RMSE value for 40TPDecay instance. Other example: minimum MAE value is achieved with ϵ_{R^2} calibration for 50TPDecay instance, while minimum ϵ_{R^2} value is achieved with MAE calibration for 50TPMedium instance. These situations also occur with MAPE: calibration with MAE achieves minimum MAPE value for 50TPMedium instance.

TABLE II: Summary Results of Calibration with Different Algorithms: Mean, Standard Deviation and Minimum of the MAPE of 20 Calibrations

		RAND	SSGA	IPOP-CMA-ES
40TP (144) MEDIUM	**Mean**	35.966	18.030	**4.030**
	Dev. Std.	0.575	2.567	**0.479**
	Min	34.931	13.302	**3.030**
50TP (174) DECAY	**Mean**	240.074	31.239	**29.690**
	Dev. Std.	17.433	**2.857**	7.748
	Min	188.787	23.369	**17.250**

In any case, none of these metrics dominates the other, and all the metrics have cases for and against. Thereby, these results do not show that there is one metric outperforming the rest.

Table IV shows the results of the calibration experiments by considering Soft-DTW measure for all the instances. We can see that, as well as with point-to-point metrics, the fitting is better in medium instances, regardless the number of touchpoints. This could indicate that Soft-DTW calibration is behaving similar to the calibration with the previous metrics. In order to obtain a more rigorous comparison, we will perform ranking statistical tests, plot the models outputs, and show the values obtained for some parameters in the calibrated models in the next sub-sections.

D. Ranking Statistical Tests

We perform ranking statistical tests [23] over the calibration results with all metrics. We apply *post-hoc* procedures (Friedman and Bonferroni-Dunn tests) to find how significant the differences between errors with different metrics are.

In particular, we apply five ranking tests. Each ranking uses the fitness value for one specific metric of the $20 \cdot 6 = 120$ calibrated models. Hence, each ranking performs 120 comparisons, and in each comparison it is assigned a rank value (1 to 5) for the calibration with each metric. Finally, each ranking computes a mean rank value for each metric.

Table V shows the ranking result for each metric along with p-values for Friedman and Bonferroni-Dunn tests. We consider a significance level of $\alpha = 0.01$ for Friedman test and of $\alpha = 0.05$ for Bonferroni-Dunn test. Since p-value for Friedman test is lower than the significance level in all cases, we conclude that there are significant differences between the calibration results for all metrics evaluating with any other metric.

Bonferroni-Dunn p-values show that only ϵ_{R^2} and MAPE calibration have significantly worse MAE results than MAE calibration, since their p-values are lower than $\alpha = 0.05$. This also occurs with RMSE evaluation. On the other hand, in the cases of MAPE, ϵ_{R^2} and Soft-DTW evaluation, calibration with any metric give significantly worse results than calibration with the evaluation metric.

We can conclude from these results that MAE and RMSE calibration, and Soft-DTW to a lesser extent, have a similar behavior. MAPE and ϵ_{R^2} present more differences with respect to the rest.

E. Calibrated Parameters and Model's Outputs

Finally, we inspect in this section the outputs and parameters of the calibrated models. In order to compare outputs, we plot the mean awareness of the 20 calibrations for each metric, instance and brand. We include here the result for brand 1 in 40TPDecay instance (Fig. 2), brand 5 in 45TPDecay instance (Fig. 3) and brand 8 in 50TPDecay (Fig. 4)

TABLE III: Summary Results (Mean, Standard Deviation. and Minimum) for Calibration with Point-to-Point Metrics

		MAE Calibration				RMSE Calibration				MAPE Calibration				R2 Calibration			
		MAE	**RMSE**	**MAPE**	**R2**	**MAE**	**RMSE**	**MAPE**	**R2**	**MAE**	**RMSE**	**MAPE**	**R2**	**MAE**	**RMSE**	**MAPE**	**R2**
40TP (144) DECAY	**Mean**	**4.485**	**5.415**	45.603	70.025	4.670	5.530	45.636	80.339	6.306	7.617	**26.315**	171.299	5.178	6.205	48.571	**49.739**
	Dev. Std.	0.797	0.930	12.417	31.140	0.478	0.562	12.853	26.395	1.147	1.428	6.064	91.020	0.788	0.854	10.750	10.338
	Min	**3.172**	**3.854**	25.183	**24.964**	3.426	4.192	25.098	48.992	4.605	5.339	**16.023**	64.467	3.462	4.323	27.425	26.101
40TP (144) MEDIUM	**Mean**	1.784	2.139	4.047	33.344	**1.734**	**2.074**	**3.899**	32.425	1.949	2.283	4.030	38.552	1.847	2.203	4.171	**30.605**
	Dev. Std.	0.184	0.203	0.459	7.009	0.224	0.252	0.502	7.891	0.314	0.314	0.479	12.643	0.197	0.215	0.471	3.995
	Min	1.419	1.725	3.217	22.371	**1.370**	**1.670**	3.078	21.361	1.394	1.693	**3.030**	**20.582**	1.480	1.784	3.228	21.916
45TP (159) DECAY	**Mean**	4.720	5.814	63.152	65.360	**4.584**	**5.676**	57.984	60.244	6.265	7.484	**29.846**	138.083	5.330	6.557	52.299	**44.761**
	Dev. Std.	0.768	0.934	16.682	25.558	0.743	0.887	16.596	20.827	1.253	1.377	7.156	93.852	0.855	1.005	11.937	13.918
	Min	2.689	3.405	35.683	22.335	**2.679**	**3.342**	23.865	**17.700**	4.026	5.080	**18.638**	47.975	3.959	4.947	33.714	25.268
45TP (159) MEDIUM	**Mean**	**2.125**	**2.524**	**4.915**	44.522	2.193	2.593	5.067	47.484	2.387	2.797	5.150	59.263	2.182	2.600	5.177	**43.194**
	Dev. Std.	0.271	0.284	0.622	10.678	0.281	0.302	0.562	12.790	0.422	0.436	0.734	23.500	0.363	0.396	0.817	13.777
	Min	1.768	2.145	3.981	30.445	**1.613**	**1.949**	**3.719**	27.328	1.718	2.090	3.874	29.901	1.699	2.095	4.067	**25.844**
50TP (174) DECAY	**Mean**	**3.529**	**4.421**	63.268	40.600	3.618	4.449	64.541	42.955	4.832	5.864	**29.690**	65.242	3.777	4.676	46.884	**29.005**
	Dev. Std.	0.534	0.667	20.641	13.771	0.451	0.530	16.626	13.780	1.242	1.408	7.748	39.788	0.812	0.989	10.499	9.892
	Min	2.851	3.581	32.898	24.249	2.870	3.500	40.804	23.325	3.157	3.976	**17.250**	22.251	**2.601**	**3.186**	27.672	**14.089**
50TP (174) MEDIUM	**Mean**	**1.739**	**2.113**	4.553	45.900	1.810	2.174	4.681	50.201	1.945	2.330	**4.505**	50.421	1.929	2.322	4.633	**44.171**
	Dev. Std.	0.133	0.167	0.405	6.475	0.166	0.180	0.502	14.799	0.240	0.249	0.458	14.803	0.150	0.161	0.334	5.548
	Min	**1.484**	**1.801**	**3.733**	**33.632**	1.593	1.925	3.754	33.982	1.580	1.919	3.769	34.915	1.675	2.023	4.043	34.697

TABLE IV: Summary Results (Mean, Dev. std. and Min) for Calibration with Soft-DTW

	40TPdec	**40TPmed**	**45TPdec**	**45TPmed**	**50TPdec**	**50TPmed**
Mean	25.892	9.205	24.215	11.290	20.146	9.189
Dev. Std.	3.997	0.996	4.047	1.594	2.749	0.986
Min	19.989	7.731	15.774	8.869	16.837	7.445

We can see that all metrics get solutions with similar trend, and although these solutions are not fully in line with the historical data, they are close to them. MAPE obtains the most different solutions, and in general they are the furthest from the objective.

It should be noted that solutions with Soft-DTW do not differ from the rest. In fact, in line with ranking tests, MAE and RMSE solutions are closer to Soft-DTW than to ϵ_{R^2} and MAPE solutions, although they are also point-to-point metrics. Consequently, it seems that considering trend evolution of time series in calibration not lead to better solutions.

We also plot the parameters of the models by boxplots to observe if calibrating with different metrics results in the exploration of different search spaces regions. We show here boxplots for buzz increment (Fig. 5) and awareness impact of touchpoint 0 (Fig. 7) in 40TPMedium instance, in addition to awareness decay (Fig. 6) and awareness impact of touchpoint 0 (Fig. 8) in 50TPDecay instance. Furthermore, we include the optimum value in the caption of each boxplot, i.e., the value used to generate the artificial historical data of the corresponding instance.

As we can see, although there are some particular cases such as the outlier calibrating with ϵ_{R^2} in Fig. 7, calibration with different metrics does not lead to significant differences on parameters. On the other hand, as well as we observed with the errors, resulting parameters for decay instance have more variability than for medium instance. Finally, although the parameter values obtained are close to the optimum value, results for awareness impact of touchpoint 0 in 40TPMedium instance are far from the optimum. Considering that calibration in medium instances obtained lower errors, this fact illustrates that a good calibration does not ensure that the values of the parameters are correct. It means that it is also needed to perform a validation phase with the help of an expert.

VI. Concluding Remarks

In this paper we performed a study of the deviation measure influence in the calibration of an agent-based model with IPOP-CMA-ES through an exhaustive experimentation. Specifically, we have consider four classic point-to-point metrics (MAE, RMSE, MAPE and ϵ_{R^2}) and a more sophisticated one based on Soft-DTW, which is capable of considering trend evolution. In addition, we applied and compared the quality of IPOP-CMA-ES for calibration. IPOP-CMA-ES obtains much better results that calibrating with a SSGA or a random search.

We completed metric comparison with ranking statistical tests and the analysis of the outputs and the parameters of the calibrated model. In the view of the results, although MAE, RMSE and Soft-DTW are the metrics which report the closest outputs to the historical data, we find a similar behavior for all of them. This is particular significant for the case of Soft-DTW, which not improves the solutions viewed with point-to-point metric. In fact, Soft-DTW shows a behavior especially similar to MAE and RMSE, and its solutions present the same trends for awareness (output considered) than the rest of metrics.

The explanation of these results might lie in the awareness dynamics of the agent-based model instances used. It is possible that the calibration parameters considered do not affect to the changes in trend of the awareness. In this way, the Soft-DTW ability of considering solutions with shifted outputs at temporal axis would not present any advantage.

We propose as future works studying the influence of calibration with metrics based on trends using other instances or models with different temporal dynamics.

Acknowledgment

This work is supported by the Spanish Agencia Estatal de Investigación, the Andalusian Government, the University of Granada, European Regional Development Funds (ERDF) under grants EXASOCO (PGC2018-101216-B-I00), SIMARK (P18-TP-4475), and AIMAR (A-TIC-284-UGR18).

References

[1] E. Bonabeau, "Agent-based modeling: Methods and techniques for simulating human systems," *Proceedings of the national academy of sciences*, vol. 99, no. suppl 3, pp. 7280–7287, 2002.

TABLE V: Calibration Ranking with respect to each Metric, Friedman Test and Bonferroni p-values

	Friedman test	Ranking					
MAE Ranking	χ^2_F = 75.393	**Rank**	MAE	RMSE	Soft-DTW	R2	MAPE
			2.44	2.50	2.86	3.23	3.97
	p-value = $1.645 \cdot 10^{-15}$	**Bonferroni** p	-	1.00	0.09	$5.9 \cdot 10^{-05}$	$2.8 \cdot 10^{-16}$
RMSE Ranking	χ^2_F = 74.74	**Rank**	RMSE	MAE	Soft-DTW	R2	MAPE
			2.48	2.53	2.78	3.23	3.98
	p-value = $2.262 \cdot 10^{-15}$	**Bonferroni** p	-	1.00	$4 \cdot 10^{-01}$	$2 \cdot 10^{-04}$	$8.9 \cdot 10^{-16}$
MAPE Ranking	χ^2_F = 55.447	**Rank**	MAPE	R2	RMSE	Soft-DTW	MAE
			2.05	3.12	3.21	3.30	3.33
	p-value = $2.619 \cdot 10^{-11}$	**Bonferroni** p	-	$2.1 \cdot 10^{-08}$	$9.2 \cdot 10^{-10}$	$3.1 \cdot 10^{-11}$	$1.2 \cdot 10^{-11}$
R2 Ranking	χ^2_F = 93.793	**Rank**	R2	MAE	RMSE	Soft-DTW	MAPE
			2.01	2.88	2.99	3.15	3.97
	p-value $< 2.2 \cdot 10^{-16}$	**Bonferroni** p	-	$6.7 \cdot 10^{-06}$	$3 \cdot 10^{-07}$	$1.7 \cdot 10^{-09}$	$< 2 \cdot 10^{-16}$
Soft-DTW Ranking	χ^2_F = 100.13	**Rank**	Soft-DTW	MAE	RMSE	R2	MAPE
			2.14	2.72	2.74	3.37	4.03
	p-value $< 2.2 \cdot 10^{-16}$	**Bonferroni** p	-	0.0066	0.0041	$8.1 \cdot 10^{-11}$	$< 2 \cdot 10^{-16}$

[2] A. J. McLane, C. Semeniuk, G. J. McDermid, and D. J. Marceau, "The role of agent-based models in wildlife ecology and management," *Ecological Modelling*, vol. 222, no. 8, pp. 1544–1556, 2011.

[3] W. Rand, J. Herrmann, B. Schein, and N. Vodopivec, "An agent-based model of urgent diffusion in social media," *Journal of Artificial Societies and Social Simulation*, vol. 18, no. 2, p. 1, 2015.

[4] S. L. Chang, N. Harding, C. Zachreson, O. M. Cliff, and M. Prokopenko, "Modelling transmission and control of the covid-19 pandemic in australia," *Nature communications*, vol. 11, no. 1, pp. 1–13, 2020.

[5] M. Chica and W. Rand, "Building agent-based decision support systems for word-of-mouth programs: a freemium application," *Journal of Marketing Research*, vol. 54, no. 5, pp. 752–767, 2017.

[6] F. Stonedahl and W. Rand, "When does simulated data match real data?" in *Advances in Computational Social Science*. Springer, 2014, pp. 297–313.

[7] M. Chica, J. Barranquero, T. Kajdanowicz, S. Damas, and Ó. Cordón, "Multimodal optimization: an effective framework for model calibration," *Information Sciences*, vol. 375, pp. 79–97, 2017.

[8] I. Moya, M. Chica, and Ó. Cordón, "A multicriteria integral framework for agent-based model calibration using evolutionary multiobjective optimization and network-based visualization," *Decision Support Systems*, vol. 124, p. 113111, 2019.

[9] J. S. Lee, T. Filatova, A. Ligmann-Zielinska, B. Hassani-Mahmooei, F. Stonedahl, I. Lorscheid, A. Voinov, G. Polhill, Z. Sun, and D. C. Parker, "The complexities of agent-based modeling output analysis," *The journal of artificial societies and social simulation*, vol. 18, no. 4, 2015.

[10] X. Li, A. Engelbrecht, and M. Epitropakis, "Results of the 2013 ieee cec competition on niching methods for multimodal optimization," in *Report presented at 2013 IEEE Congress on Evolutionary Computation Competition on: Niching Methods for Multimodal Optimization*, 2013.

[11] I. Moya, E. Bermejo, M. Chica, and Ó. Cordón, "Coral reefs optimization algorithms for agent-based model calibration," *Engineering Applications of Artificial Intelligence*, vol. 100, p. 104170, 2021.

[12] M. Cuturi and M. Blondel, "Soft-dtw: a differentiable loss function for time-series," in *International Conference on Machine Learning*. PMLR, 2017, pp. 894–903.

[13] E. Keogh and C. A. Ratanamahatana, "Exact indexing of dynamic time warping," *Knowledge and information systems*, vol. 7, no. 3, pp. 358–386, 2005.

[14] G. Ten Broeke, G. Van Voorn, and A. Ligtenberg, "Which sensitivity analysis method should i use for my agent-based model?" *Journal of Artificial Societies and Social Simulation*, vol. 19, no. 1, 2016.

[15] J. Zhong and W. Cai, "Differential evolution with sensitivity analysis and the powell's method for crowd model calibration," *Journal of computational science*, vol. 9, pp. 26–32, 2015.

[16] E. C. T. Zúñiga, I. L. L. Cruz, and A. R. García, "Parameter estimation for crop growth model using evolutionary and bio-inspired algorithms," *Applied Soft Computing*, vol. 23, pp. 474–482, 2014.

[17] D. Ngoduy and M. Maher, "Calibration of second order traffic models using continuous cross entropy method," *Transportation Research Part C: Emerging Technologies*, vol. 24, pp. 102–121, 2012.

[18] C. Dai, M. Yao, Z. Xie, C. Chen, and J. Liu, "Parameter optimization for growth model of greenhouse crop using genetic algorithms," *Applied Soft Computing*, vol. 9, no. 1, pp. 13–19, 2009.

[19] D. Muraro and R. Dilão, "A parallel multi-objective optimization algorithm for the calibration of mathematical models," *Swarm and Evolutionary computation*, vol. 8, pp. 13–25, 2013.

[20] A. E. Eiben and J. E. Smith, "What is an evolutionary algorithm?" in *Introduction to Evolutionary Computing*. Springer, 2015, pp. 25–48.

[21] N. Hansen and S. Kern, "Evaluating the CMA evolution strategy on multimodal test functions," in *Parallel Problem Solving from Nature - PPSN VIII, 8th International Conference, Birmingham, UK, September 18-22, 2004, Proceedings*, ser. Lecture Notes in Computer Science, X. Yao, E. K. Burke, J. A. Lozano, J. Smith, J. J. M. Guervós, J. A. Bullinaria, J. E. Rowe, P. Tiño, A. Kabán, and H. Schwefel, Eds., vol. 3242. Springer, 2004, pp. 282–291. [Online]. Available: https://doi.org/10.1007/978-3-540-30217-9_29

[22] A. Auger and N. Hansen, "A restart cma evolution strategy with increasing population size," in *2005 IEEE congress on evolutionary computation*, vol. 2. IEEE, 2005, pp. 1769–1776.

[23] J. Derrac, S. García, D. Molina, and F. Herrera, "A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms," *Swarm and Evolutionary Computation*, vol. 1, no. 1, pp. 3–18, 2011.

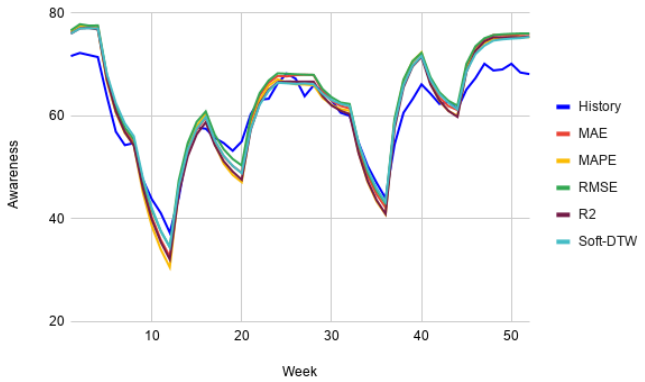

Fig. 2: Awareness Evolution Brand 1 in 40TPDecay

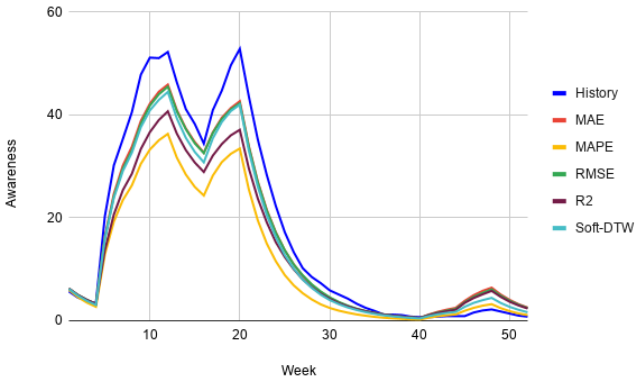

Fig. 3: Awareness Evolution Brand 5 in 45TPDecay

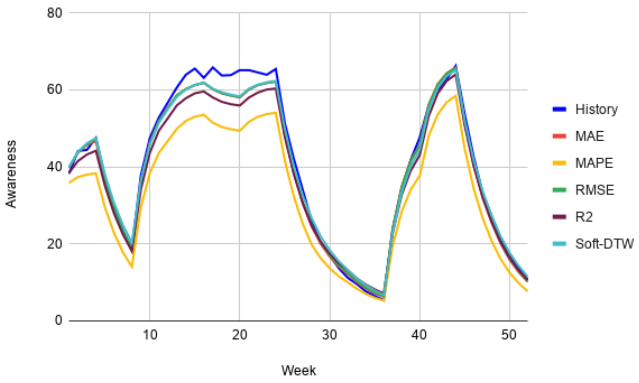

Fig. 4: Awareness Evolution Brand 8 in 50TPDecay

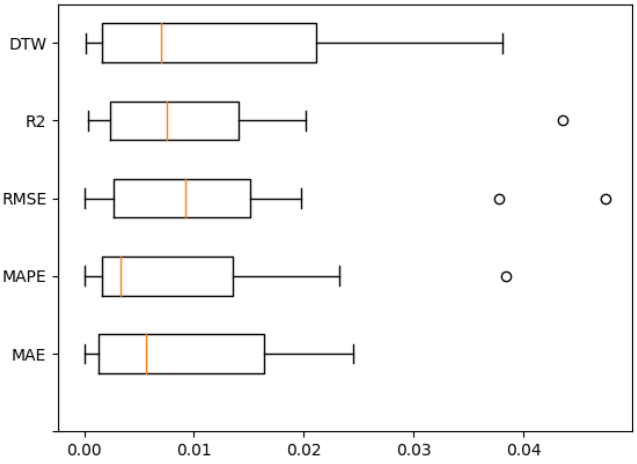

Fig. 5: Sensitivity Buzz Increment TP 0, 40TPMedium (optimum value of 0.2)

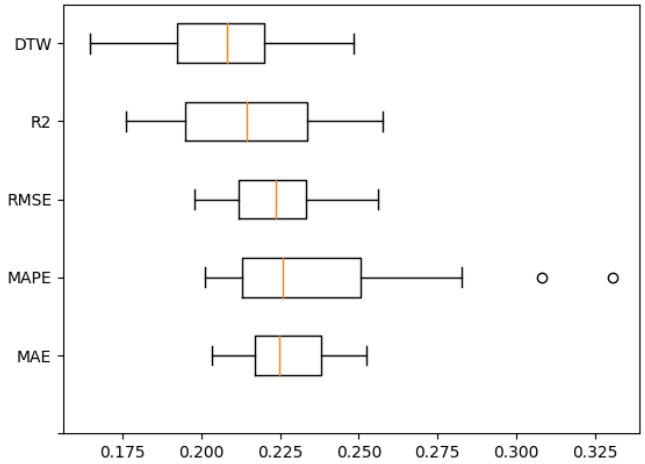

Fig. 6: Sensitivity Awareness Decay, 50TPDecay (optimum value of 0.2)

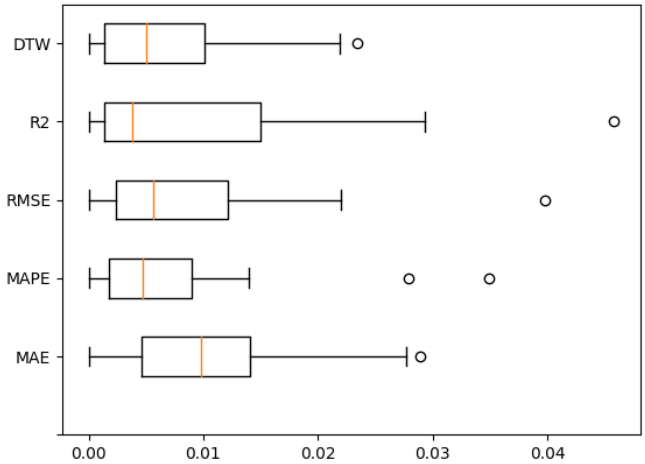

Fig. 7: Sensitivity Awareness Impact TP0, 40TPMedium (optimum value of 0.009)

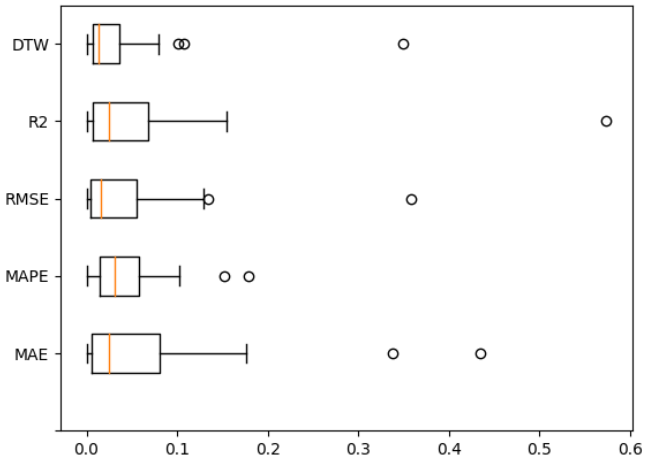

Fig. 8: Sensitivity Awareness Impact TP0, 50TPDecay (optimum value of 0.001)