Statistical and Economic Evaluation of Time Series Models for Forecasting Arrivals at Call Centers

Bastianin, Andrea and Galeotti, Marzio and Manera, Matteo

December 2016

Statistical and Economic Evaluation of Time Series Models for Forecasting Arrivals at Call Centers

Andrea Bastianin
University of Milan, Italy

Marzio Galeotti
University of Milan
IEFE Bocconi, Italy

Matteo Manera
University of Milan-Bicocca, Italy
Fondazione Eni Enrico Mattei, Milan

December, 2016

Abstract: Call centers' managers are interested in obtaining accurate forecasts of call arrivals because these are a key input in staffing and scheduling decisions. Therefore their ability to achieve an optimal balance between service quality and operating costs ultimately hinges on forecast accuracy. We present a strategy to model selection in call centers which is based on three pillars: (*i*) a flexible loss function; (*ii*) statistical evaluation of forecast accuracy; (*iii*) economic evaluation of forecast performance using money metrics. We implement fourteen time series models and seven forecast combination schemes on three series of call arrivals. We show that second moment modeling is important when forecasting call arrivals. From the point of view of a call center manager, our results indicate that outsourcing the development of a forecasting model worth its cost, since the simple Seasonal Random Walk model is always outperformed by other, relatively more sophisticated, specifications.

Key Words: ARIMA; Call center arrivals; Loss function; Seasonality; Telecommunications forecasting.

JEL Codes: C22, C25, C53, D81, M15.

Corresponding author: Andrea Bastianin, University of Milan, Department of Economics, Management and Quantitative Methods,Via Conservatorio 7, 20122 Milan, Italy. Email: andrea.bastianin@unimi.it.

1 Introduction

Hiring, staffing and scheduling are strategic decisions for the management of call centers, that represent a highly labor-intensive and large services industry, in which human resources costs account for 60-70% of the operating budget (Gans et al., 2003). Forecasts of call arrivals are a key input for choices relating to the acquisition and deployment of human resources, therefore they ultimately determine the ability of managers to achieve an optimal balance between service quality and operating costs (Akşin et al., 2007).

We present a novel strategy to select time series models of call arrivals that is based on three pillars: (*i*) a flexible loss function; (*ii*) statistical evaluation of forecast accuracy; (*iii*) economic evaluation of forecast performance using money metrics.

The use of a flexible loss function, as well as the implementation of statistical tests to rank and select forecasting models represent the first novelty of this paper. In fact, in this strand of the literature, most studies only provide model rankings based on symmetric loss functions or informal forecast comparisons (see Bastianin et al., 2011 and Ibrahim et al., 2016 for a survey). Since over-forecasting leads to over-staffing and hence unnecessarily high operating costs, while under-forecasting results in under-staffing and hence low service quality, the choice of the metric used to evaluate competing time series models depends on the preferences of the call center management, that are not necessarily well approximated by a symmetric loss function.

An asymmetric loss function is required when the call center operates under a Service Level Agreement (SLA). An 80/20 SLA is widely applied and implies that eighty percent of the incoming calls must be answered within twenty seconds (Stolletz, 2003). In this case, a reasonable degree of over-forecasting is less costly than the same amount of under-prediction, possibly because agents becoming free at short notice might be assigned to meetings and training (Taylor, 2008).

We rely on the loss function put forth by Elliott et al. (2005) that nests both symmetric and asymmetric loss functions as special cases and hence describes a wide range of call center managers' preferences.

The second novelty of this study is the translation of statistical measures of forecasting

performance into money metrics. Money measures of performance, are intimately related to the profit maximizing behavior of economic agents and hence should be considered by call center managers as more intuitive evaluation instruments to complement loss functions and statistical tests (Leitch and Tanner, 1991).

We estimate fourteen time series models — including the Seasonal Random Walk (SRW) as a benchmark — that capture different key features of daily call arrivals. These data are characterized by the presence of intra-weekly and intra-yearly seasonality, inter-day dependency (i.e. non-zero auto-correlation), overdispersion (i.e. the variance of the arrival count per time period is larger than its expected value) and conditional heteroskedasticity (see e.g. Ibrahim et al., 2016). These features are shared by the three series we analyze, that are daily arrivals at call centers operated by an Italian electric utility and by two retail banks, one in the U.S. and the other in Israel.

Moreover, since it is well documented that combined forecasts often outperform forecasts generated by individual models (Timmermann, 2006), the third novelty of the paper is to implement seven forecast combination methods applied to five sets of models. Overall, we produce a total of 47 alternative forecasts.

We show that second moment modeling using Generalized Autoregressive Conditional Heteroskedasticity (GARCH) models is the preferred approach when forecasting daily call arrivals, thus suggesting that volatility modeling is useful. This result holds true not only for the call center managed by the Italian utility, but also for the two retail banks' arrival series.

From the point of view of a call center manager, our results indicate that outsourcing the development of forecasting models could be worth its cost: the simple SRW model is always outperformed by other, relatively more sophisticated, but easily implementable, specifications.

The economic evaluation of forecast accuracy is somehow similar to that of Shen and Huang (2008), however there are some differences between the two approaches. First, the forecast horizon is different; second, their accuracy metric is the staffing level, while ours is the money the manager can earn; third, they rely on the Root Mean Squared that is a symmetric loss function, while we design a compensation scheme which penalizes under-

staffing more heavily than over-staffing.

The paper of Taylor (2008) is related to our work because, to the best of our knowledge, is the only paper that evaluates a large number of time series models and combination schemes. Compared with Taylor (2008), we do not deal with intra-day forecasts, but we evaluate the performance of models more thoroughly.

There are at least two ways in which our results can be used to derive intra-daily forecasts. First, we can think that daily arrivals are used in a top-down approach where they are disaggregated to high frequency by some procedure (Gans et al., 2003). Alternatively, we can assume that the manager of the call center designs intra-day staffing schedules on the basis of a judgmental process tha is based on its forecasts of daily totals.

The plan of the paper is as follows. Section 2 describes data, empirical methods and our approach to the economic evaluation of forecasts. Empirical results are presented and discussed in Section 3, while Section 4 concludes.

2 Data and methods

2.1 Data

We have collected three series recording the daily call arrivals received by call centers in different industries and countries. The main results of the paper are based on call arrivals for a call center managed by an anonymous Italian energy utility. Robustness checks in Section 3.3 replicate the analysis for two alternative series that are call arrivals at call centers operated by two anonymous retail banks, one located in Israel and the other in the U.S..[1]

Call arrivals at the call center operated by the Italian energy utility range from September 2008 until September 2010, for a total of $T = 749$ observations. The call center, whose employees help customers with invoicing problems, operates fourteen hours per day and is closed only during public holidays. Closing days are known in advance and hence are kept

[1] A plot of the Italian data appears in Bastianin et al. (2011). We thank Avi Mandelbaum for providing access to the data for retail banks through the Technion Service Enterprise Engineering (SSE) Laboratory. See `http://ie.technion.ac.il/Labs/Serveng` for a detailed description of these series.

in the estimation sample by substituting zero observations with the number of calls recorded during the previous week; forecasts for these days are subsequently set to zero.

This series exhibits all the features that are typical of call arrivals data. It displays both daily and monthly seasonality: the number of incoming calls decreases steadily from Monday to Sunday; moreover, given the nature of the service provided by the company, the intensity of calls varies with the season of the year, peaking during winter and summer. Lastly, even controlling for seasonality and autoregressive dynamics, the LM test for ARCH effects rejects the null hypothesis of homoskedasticity.

The call arrivals for the two banks have similar characteristics: seasonality, as well as conditional heteroskedasticity.

2.2 Forecasting models and methods

Given the nature of our series, all models have been chosen so as capture different key features such as the presence of autocorrelation, seasonality, overdispersion and conditional heteroskedasticity. The empirical specifications implemented in the analysis are shown in Table 1 and can be ideally assigned to three groups.

The first group includes: the Seasonal Random Walk (SRW) model, which is used as a benchmark, and a variety of time series models. We focus on well established specifications such as: the Box-Jenkins Airline model, ARMAX and SARMAX models with and without GARCH effects, the Periodic Autoregressive (PAR) model and Holt-Winters exponential smoothing.

The total number of calls arriving at call centers in a given time period is a count and as such it is usually modeled as a Poisson arrival process (Gans et al., 2003). The underlying assumption is that there is large population of potential customers, each of which makes calls independently with a very low probability. Simple dynamic models for count data are then further plausible specifications.

Jung and Tremayne (2011) have shown that when forecasting with count data there is not a dominating modeling approach, therefore we consider three specifications based on the Exponential, Poisson and Negative Binomial distribution, respectively. While the Poisson

Table 1: Summary of models

id	Name	Dependent Variable	Explanatory Variables
$\mathcal{M}_0$	Seasonal Random Walk	Y_t	Y_{t-7}
$\mathcal{M}_1$	ARMAX	y_t	AR(1), AR(7), AR(8), MA(1), $\mathbf{D}_t$
$\mathcal{M}_2$	ARMAX-GARCH(1,1)	y_t	AR(1), AR(7), AR(8), MA(1), $\mathbf{D}_t$
$\mathcal{M}_3$	TVD-AR	y_t	$\mathbf{D}_t$
$\mathcal{M}_4$	SARMAX	y_t	AR(1), SAR(7), MA(1), SMA(28), $\mathbf{D}_t$
$\mathcal{M}_5$	SARMAX-GARCH(1,1)	y_t	AR(1), SAR(7), MA(1), SMA(28), $\mathbf{D}_t$
$\mathcal{M}_6$	PAR(2)	y_t	$y_{t-1}, y_{t-2}, \mathbf{D}_t$
$\mathcal{M}_7$	Airline	$\Delta \times \Delta_7 y_t$	MA(1), SMA(8)
$\mathcal{M}_8$	Poisson	Y_t	$Y_{t-1}, \mathbf{D}_t$
$\mathcal{M}_9$	NegBin	Y_t	$Y_{t-1}, \mathbf{D}_t$
$\mathcal{M}_{10}$	Exponential	Y_t	$Y_{t-1}, \mathbf{D}_t$
$\mathcal{M}_{11}$	MEM	$y_t/\hat{y}_{SR,t}$	$y_{t-1}/\hat{y}_{SR,t-1}, \mathbf{D}_t$
$\mathcal{M}_{12}$	Spline-SARX	$y_t/\hat{y}_{LR,t}$	AR(1), SAR(7), $\mathbf{D}_t$
$\mathcal{M}_{13}$	Holt-Winters	y_t	Multiplicative

Notes: Y_t is the number of incoming calls; $y_t \equiv \log Y_t$; $\mathbf{D}_t$ is a vector of dummies, one for each day of the week; $\Delta_k = (1 - L^k)$ where L is the lag operator. $\hat{y}_{SR,t}$ denotes fitted values from the regression of y_t on the vector of dummies; $\hat{y}_{LR,t}$ denotes fitted values from the interpolation of y_t with a natural cubic spline, with the number of knots that equals the number of months in the sample; ARMA and seasonal ARMA terms are denoted as AR(.), MA(.), SAR(.) and SMA(.), where the number in brackets represents their order; "multiplicative" indicates that forecasts from $\mathcal{M}_{13}$ are obtained with the multiplicative Holt-Winters exponential smoothing, see Gardner (2006) for details.

model requires the variance of the arrival count per time period to be equal to its expected value, the Negative Binomial relaxes this assumption.

We add to this set of standard models, a third group of seasonal autoregressive specifications that have not been previously used to predict incoming calls (see Taylor, 2008, and references therein). These include: a linear model with smoothly changing deterministic seasonality (i.e. the Time Varying Dummy AR, TVD-AR, model of Franses and van Dijk, 2005) and the Multiplicative Error Model (MEM) of Engle (2002). Lastly, we also forecast with a SARMAX model applied to the series of incoming calls, after monthly seasonality has been removed with a natural cubic spline function.

Estimation and forecasting of models is carried out recursively: the estimation sample expands by including a new observation at each iteration.[2] The first iteration relies on an estimation sample of $R = 371$ days. The forecast horizon, h, ranges from one day to one month, that is $h = 28$ days. The recursive scheme implies that the number of predictions,

[2]Although the recursive or expanding scheme has the advantage of using more observations than the rolling forecasting scheme, the latter has the is robust to the presence of structural breaks. However, for our series there is no evidence of structural breaks.

P_h for $h = 1, \ldots, 28$, varies from $P_1 = 378$ to $P_{28} = 351$.

Model selection is performed only once using the sample of data pertaining to the first iteration of the estimation-forecasting scheme. We have selected most of the specifications in Table 1 using both the Schwarz Information Criterion (SIC) to choose the optimal number of lags (or, for some models, SARMA terms) and a stepwise regression approach.[3] Models are also subject to passing a Lagrange Multiplier test for first-to-eighth order residual autocorrelation at the 5% confidence level. For ARMA(p,q) models we set $p^{\max}, q^{\max} = 28$; for seasonal AR(k) and MA(l) terms, we tried the following $k^{\max}, l^{\max} = 7, 14, 21, 28$.

Some models include a GARCH component because squared residuals from ARMAX and SARMAX specifications display some un-modeled dynamics. Moreover, the inclusion of a GARCH equation can help us to shed light on the usefulness of second moment modeling for forecasting call arrivals.

2.3 Combining methods

Given that combined forecasts are often found to outperform individual models (see Timmermann, 2006), we implement many combination schemes to predict call arrivals.

As shown in Table 2, our fourteen models are collected into five groups that always exclude the benchmark SRW model. The first set, $\mathcal{G}_1$, excludes only the Holt-Winters exponential smoothing. The second group of models, $\mathcal{G}_2$, includes two ARMAX specifications and the TVD-AR model. Group $\mathcal{G}_3$ differs from $\mathcal{G}_4$ in that the latter excludes the PAR model, from the set containing ARMAX, TVD-AR and SARMAX models. The last group, $\mathcal{G}_5$, is made up of models for time series of count data.

We combine forecasts from these five groups of models with average, trimmed average, median, minimum, maximum and Approximate Bayesian Model Averaging (ABMA) combining schemes, see Table 3. All these methods have a feature in common: they do not require holding out a set of out-of-sample observations and hence they can be used in real-time by the forecast user.

Stock and Watson (2004) have shown that simple combining methods such as the av-

[3]The stepwise approach starts with a model including all terms selected with the SIC; then it is repeatedly applied until all variables are significant at the 5% critical level.

Table 2: Groups of models for combining methods

id	models
$\mathcal{G}_1$	$\mathcal{M}_i$ for $i = 1, ..., 12$
$\mathcal{G}_2$	$\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3$
$\mathcal{G}_3$	$\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3, \mathcal{M}_4, \mathcal{M}_5, \mathcal{M}_6$
$\mathcal{G}_4$	$\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3, \mathcal{M}_4, \mathcal{M}_5$
$\mathcal{G}_5$	$\mathcal{M}_8, \mathcal{M}_9, \mathcal{M}_{10}$

Notes: see Table 1.

erage, the trimmed average and the median work well in macroeconomic forecasting. If compared to the simple average, both the median forecast and the trimmed average combination method (that excludes the highest and the lowest forecasts) reduce the impact of individual outlying forecasts. The maximum and minimum combination methods are used to represent two opposite situations, in which a manager is either adverse to under-staffing, or is trying to minimize labor costs and is not subject to any kind of SLA.

Table 3: Combining methods

Method	Description
Average (c_1)	$f^{c_1}_{\mathcal{G}_i,t} = \frac{1}{M_{\mathcal{G}_i}} \sum_{m=1}^{M_{\mathcal{G}_i}} f_{m,t}$
Trimmed Average (c_2)	$f^{c_2}_{\mathcal{G}_i,t} = \frac{1}{M_{\mathcal{G}_i}-2} \sum_{m=1}^{\left(M_{\mathcal{G}_i}-2\right)} f_{m,t}$
Median (c_3)	$f^{c_3}_{\mathcal{G}_i,t} = \text{median}\left(\mathbf{f}_{\mathcal{G}_i,t}\right)$
Min (c_4)	$f^{c_4}_{\mathcal{G}_i,t} = \max\left(\mathbf{f}_{\mathcal{G}_i,t}\right)$
Max (c_5)	$f^{c_5}_{\mathcal{G}_i,t} = \min\left(\mathbf{f}_{\mathcal{G}_i,t}\right)$
ABMA-SIC (c_6)	$f^{c_6}_{\mathcal{G}_i,t} = \sum_{m=1}^{M_{\mathcal{G}_i}} w^{c_j}_{m,t} f_{m,t}$
ABMA-AIC (c_7)	$f^{c_7}_{\mathcal{G}_i,t} = \sum_{m=1}^{M_{\mathcal{G}_i}} w^{c_j}_{m,t} f_{m,t}$

Notes: $f^{c_j}_{\mathcal{G}_i,t}$ denotes the forecast at time t obtained with combining method j on $\mathcal{G}_i$, for $j = 1, \ldots, 7$ and $i = 1, \ldots, 5$; $f_{m,t}$ is the forecast at time t from model m, for $m = 1, \ldots, M_{\mathcal{G}_i}$, where $M_{\mathcal{G}_i}$ is the number of models in the i-th group; $\mathbf{f}_{\mathcal{G}_i,t}$ is a $\left(M_{\mathcal{G}_i} \times 1\right)$ vector of forecasts from models in $\mathcal{G}_i$; Approximate Bayesian Model Averaging (ABMA) uses weights, $w^{c_j}_{m,t} = \frac{\exp\{\zeta_{m,t}\}}{\sum_{m=1}^{M_{\mathcal{G}_i}} \exp\{\zeta_{m,t}\}}$, where $\zeta_{m,t} = \text{IC}_{m,t} - \max\left(\mathbf{IC}_{\mathcal{G}_i,t}\right)$, for $m = 1, ..., M_{\mathcal{G}_i}$, $j = 6, 7$, $i = 2, ..., 5$ and IC = SIC, AIC. The $(\mathcal{M}_{\mathcal{G}_i} \times 1)$ vector, $\mathbf{IC}_{\mathcal{G}_i,t}$, contains the IC of models in the i-th group. Combining method c_6 is based on the SIC, while c_7 uses the AIC; both exclude $\mathcal{G}_1$ from ABMA.

ABMA, successfully applied to macroeconomic forecasting by Garratt et al. (2003), uses the Schwarz and Akaike Information Criterion (SIC and AIC, respectively), to approximate the posterior probability of individual models. ABMA is applied only to models sharing the dependent variable expressed with a common unit of measure, therefore $\mathcal{G}_1$ is excluded.

2.4 Statistical measures of forecast accuracy

The need for a loss function that can be either symmetric or asymmetric arises in many economic and management problems. In the case of a call center outsourcing the production of forecasts, this family of loss functions serves two purposes: first, it allows to assign a different cost to positive and negative forecast errors; second, it helps the call center manager and the professional forecaster to decide the shape of the loss and hence to use the same metric of predictive performance.

We use $f_{i,t}$ to denote either an individual, or a combined forecast; the corresponding forecast error is $u_{i,t}$, while $\ell_{i,t}(u_{i,t})$ denotes the loss function. If not needed, we drop both model and time subscripts. Following Elliott et al. (2005), we can write:

$$\ell\left(u;\rho,\phi\right)=\left[\phi+\left(1-2\phi\right)I\left(u<0\right)\right]\left|u\right|^{\rho} \tag{1}$$

where $I(.)$ is the indicator function. The shape of the function is determined by two parameters $\rho>0$ and $\phi\in(0,1)$. The loss function is asymmetric for $\phi\neq 0.5$: over-forecasting is costlier than under-forecasting for $\phi<0.5$; on the contrary, when $\phi>0.5$, positive forecast errors (under-prediction) are more heavily weighed than negative forecast errors (over-prediction). Special cases of the function include: the quad-quad loss for $\rho=2$ and the lin-lin loss for $\rho=1$. Moreover, we get the mean absolute error (MAE) loss function for $\rho=1$ and $\phi=0.5$, and the mean square error (MSE) loss function for $\rho=2$ and $\phi=0.5$.

For each model and combination method, we produce a total of 28 series, one for each forecast horizon, h. Given that presenting detailed results for each h is not a viable option, we need a method to rank models according their overall performance. A multivariate generalization of the flexible loss function of Elliott et al. (2005) has been developed by Komunjer and Owyang (2012). Like its univariate counterpart, this function is defined by two parameters: $\rho\geq 1$ and τ, where $-1\leq\tau\leq 1$ is an asymmetry parameter, which is linked to the parameter ϕ as follows: $\tau=2\phi-1$. Let $\mathbf{u}_{h,p}$ be the p-th column of the $(H\times P)$ forecast errors matrix, then the multivariate flexible loss function can be written as:

$$\mathcal{L}\left(\mathbf{u};\rho,\tau\right)=\Big(\left\|\mathbf{u}\right\|_{\rho}+\tau\mathbf{u}\Big)\left\|\mathbf{u}\right\|_{\rho}^{\rho-1} \tag{2}$$

where $\|\mathbf{u}\|_\rho = \left(\sum_{h=1}^{H} |u_h|^\rho\right)^{1/\rho}$ is the l_p-norm.

When there is only one forecast error series, Eq. (2) reduces to $\mathcal{L}(u;\rho,\tau) = (|u| + \tau u)\,|u|^{\rho-1} = 2\,[1 - \phi + \tau I(u > 0)]\,|u|^\rho$. Notice that this expression is equivalent, up to a scale factor of two, to Equation (1). As in the univariate case, the multivariate loss includes some special cases: when $\phi = 0.5$ ($\tau = 0$) and $\rho = 2$, we obtain the trace of the MSE loss function, while for $\phi = 0.5$ ($\tau = 0$) and $\rho = 1$, Equation (2) reduces to the trace of the MAE loss function (see Zeng and Swanson, 1998). In both cases, symmetry also ensures the multivariate loss to be additively separable in univariate losses. On the contrary, when $\phi \neq 0.5$ ($\tau \neq 0$) the loss function becomes asymmetric and is not additively separable in individual losses.

2.5 Monetary measures of forecast accuracy

Dorfman and McIntosh (1997) and Leitch and Tanner (1991) have shown that money metrics of performance, such as the value of information and certainty equivalent, are more closely related to forecast's profit than traditional summary statistics based on loss functions.

Assumptions. We assume that each day t the manager uses his forecast of inbound calls for $t+1$ in an algorithm that determines the number of agents (n_t) needed to comply with the company's SLA. We impose an 80/20 SLA, implying that at least eighty percent of incoming calls should be answered within twenty seconds. The algorithm used to staff the call center is the Erlang-C queuing model: we assume that the average call duration is three minutes and that the call center is open fourteen hours a day. Notwithstanding its limitations, the Erlang-C model is widely used in practice, possibly because of its simplicity (Akşin et al., 2007; Gans et al., 2003).

Let the manager's daily payoff, W_t, be the sum of a fixed F and a variable part, v_t, that is $W_t = F + v_t(d_t)$. For the fixed part of the payoff, we impose that the call center manager earns on average 1200 Euro for 28 working days, that is $F = \lceil 1200/28 \rceil = 43$ Euro per day. The variable part of the payoff, $v_t(.)$, depends on the manager's ability to staff the call center, d_t, which is evaluated ex-post and is defined as a function of the distance between his decision, n_t, and the optimal number of agents n_t^* (i.e. n_t^* is calculated using the realized

number of incoming calls as input to the Erlang-C model).

The company relies on the compensation scheme displayed in Table 4. We have designed it so as to penalize under–staffing more heavily than over–staffing; this can be justified by assuming that the company's objective is to maximize customer satisfaction. Moreover, we assume that the company's compensation policy implies symmetry of over– and under–staffing if forecast errors of both signs exceed a certain threshold. At the end of the forecasting sample, whose length is P, the manager's payoff will be: $\pi = P \times F + \sum_{t=1}^{P} v_t$.

Following Dorfman and McIntosh (1997), we assume that the manager has a negative exponential utility function: $U(\pi) = 1 - \exp(-\lambda\pi)$, where λ represents the manager's absolute risk aversion coefficient. Notice that, for the negative exponential utility function, λ^{-1} describes the willingness to lose. Depending on d_t, the manager can either get a bonus (b_t), or be subject to a maximum penalty (p_t) of 10 Euro. Given that each day he can lose at most 20 Euro, λ is varied across the following set of values: $\lambda = [j \times P \times (b_t + p_t)]^{-1}$, where $j = 0.1, 0.5, 0.7$ denotes a percentage of the variable part of the payoff. This implies that the willingness to lose can take on the following values $\lambda^{-1} = \{732, 3660, 5124\}$ Euro. If the manager could always get the bonus, the total payoff would be $P \times (F + 10) = 18603$ Euro, where $F = 43$ Euro and $P = 351$ days; therefore the values that the willingness to lose can take on are equivalent to 0.4% 19.7% and 27.5% of the total payoff.

The end-of-period expected utility is $EU(\pi) = 1 - M_{\pi}(-\lambda)$, where $M_{\pi}(-\lambda)$ is the Moment Generating Function, MGF (see Collender and Chalfant, 1986; Elbasha, 2005; Gbur and Collins, 1989). This result and our compensation scheme allow to calculate the expected utility using Maximum Likelihood estimates of the multinomial MGF.[4]

The economic value of information. The economic value of information of a set of a forecast can only be determined with reference to the informativeness of an alternative set of forecasts. Following Dorfman and McIntosh (1997), we define the value of perfect information as the value of a model that generates perfect forecasts: $n_t = n_t^*, \ \forall t$.

If the manager could purchase this model, she would face no risk and the payoff distri-

[4]The MGF of a multinomially distributed random variable is: $M_X(t) = \left(\sum_{k=1}^{r} p_k e^{t_k}\right)^P$. An estimate of the probability p_k can be calculated as: $\hat{p}_k = \sum_{t=1}^{H} I(n_t \in CI_k)/H$, where CI_k for $k = 1, ..., 7$, denotes the naive confidence interval in Table 4.

Table 4: Multinomial payoff scheme

k	lower bound (LB_k)	upper bound (UB_k)	bonus / penalty (Euro)
1	0.00	0.80	-10
2	0.80	0.90	-5
3	0.90	0.95	-2.5
4	0.95	1.05	10
5	1.05	1.10	-1.25
6	1.10	1.20	-2.5
7	1.20	∞	-10

Notes: the multinomial compensation scheme implies that at time t the manager gets a bonus $b_t = 10$ Euro if $n_t^* \times LB_4 \leq n_t < n_t^* \times UB_4$.

bution would be a single point at $\pi^* = \max(\pi)$. The lack of risk (i.e. $var(\pi^*) = 0$) implies that the value of perfect information, V^*, is simply the payoff obtainable from the perfect forecast, in other words: $V^* = \pi^*$. Given that a forecast can be "consumed" only in discrete quantities, the expected marginal utility of the forecast, MU, equals its expected utility. We know that in equilibrium the price ratio of two goods is equal to their marginal rate of substitution, hence the value of forecasting method i is the solution of: $V_i/V^* = MU_i/MU^*$. Solving for V_i, using $MU_i = EU_i$ and $V^* = \pi^*$, yields (see Dorfman and McIntosh, 1997):

$$V_i = \frac{\pi^* EU_i}{EU(\pi^*)} \tag{3}$$

Equation (3) can be used to define the incremental value of information of forecast i with respect to model $\mathcal{M}_0$ as:

$$\Delta V_i \equiv V_i - V_{\mathcal{M}_0} \tag{4}$$

Therefore, $\mathcal{M}_i \succ \mathcal{M}_0$, if $\Delta V_i > 0$ or, equivalently, if $V_i > V_{\mathcal{M}_0}$.

The certainty equivalent. An alternative money metrics of forecast accuracy is the certainty equivalent (CE), which is defined as the value $\widetilde{\pi} \equiv CE_i$ that solves $U(\widetilde{\pi}) = EU(\pi_i)$:

$$CE_i = -\frac{1}{\lambda} \log\left[1 - EU\left(\pi_i\right)\right] \tag{5}$$

We can state that $\mathcal{M}_i \succ \mathcal{M}_j$ if $CE_i > CE_j$. The CE can also be used to determine the maximum amount of money the manager is willing to pay in order to switch from model i to model j.

We assume that the manager can choose between using the naive SRW forecast ($\mathcal{M}_0$) for free, or buying model i from an expert. Moreover, let us assume that buying forecast i costs δ_i, where δ_i represents a fraction of the payoff the manager would get from the naive forecast, that is $\delta_i \equiv \theta\pi_{\mathcal{M}_0}$ with $0 < \theta < 1$. The fraction of payoff deriving from the naive model that the manager is willing to pay to use forecast i can be written as:

$$\delta_i = CE_i - CE_{\mathcal{M}_0} \tag{6}$$

or, equivalently, as: $\theta = (CE_i - CE_{\mathcal{M}_0})/\pi_{\mathcal{M}_0}$.

2.6 Linking monetary and statistical measures of forecast accuracy

For the sake of completeness, we present a wide set of results involving symmetric and asymmetric loss functions, as well as the willingness to pay, δ_i, and the incremental value of information, ΔV_i.

More precisely, we assume the loss function in Equation (2) to be of the quad-quad kind ($\rho = 2$), while we vary the asymmetry parameter across the following set of values: $\phi = \{0.42, 0.50, 0.58\}$ (i.e. $\tau = \{0.16, 0.00, -0.16\}$). When $\phi = 0.5$, the ranking is equivalent to MSE ranking in the univariate case and coincides with the trace of MSE ranking in the multivariate case. When $\phi = 0.42$, over-forecasting is costlier than under-forecasting, and vice versa for $\phi = 0.52$.

Subsequently, in order to mimic the interaction between a professional forecaster and the call center's manager, we focus on a subset of results. At this stage, we do not see the magnitude of the asymmetry as relevant (i.e. it would not be realistic to state that a practitioner has clear opinion on the parameters of the loss function). On the contrary, we believe that is fundamental to take a stance on the direction of the asymmetry. We thus assume that the call center operates under a 80/20 SLA and is subject to a fee when the waiting time exceeds a given threshold; therefore a reasonable degree of over-staffing is less costly than the same amount of under-staffing. The presence of a fee is not the only reason

why we assume this shape for the loss function. Actually, over-staffing might be less costly than under-staffing, because agents in excess, not answering to calls, can be used for other duties within the call center.

Given this set of assumptions, when the statistical loss function is parametrized so as to penalize under-prediction more heavily than over-prediction (i.e. $\rho = 2$, $\phi = 0.52$), our economic evaluation of forecasts becomes a suitable alternative (to standard statistical metrics) to present the results to the manager of the call center.

3 Empirical results

We start this section by highlighting and discussing our main results:

- Outsourcing the development of a forecasting model worth its cost: the benchmark (SRW) model is outperformed by relatively more sophisticated, but easily implementable, methods.
- Second-moment modeling is useful: the addition of a GARCH component to ARMAX and SARMAX models improves their performance. Moreover, the SARMAX–GARCH model is among the best performing specifications.
- Combined forecasts often outperform models.
- ABMA is the preferred combining method.
- The economic and statistical evaluation of models and combining methods deliver consistent results. The best individual forecasts are those involving second-moment modeling, while the best combining method is ABMA based on a group of models that includes models with a GARCH component.
- It is reasonable to assume that the ABMA forecasting method would have higher maintenance costs than an individual model. Then, when over–forecasting is less heavily penalized than under–forecasting, the economic evaluation identifies the SARMAX–GARCH model as the best option.

Our empirical exercise mimics the interaction between a professional forecaster and the management of a call center. We assume that there are three main steps in this interaction. First, a preliminary step in which the customer and the adviser agree on the shape the loss function. The second step involves only the adviser, who uses statistical tests and loss functions to shrink the number of models to be presented to the customer. Finally, in the third step, the manager of the call center selects a forecasting method on the basis of statistical or economic criteria.

3.1 Ranking and statistical tests

Model rankings using univariate and multivariate flexible losses are presented in Table 5. The univariate loss rankings for forecast horizons of one day ($h = 1$ day), one week ($h = 7$ days) and one month ($h = 28$ days) are shown in columns 2-7, while rankings based on the multivariate loss for $h = 1, 2, \ldots, 28$ are presented in the last three columns. The analysis of forecast accuracy over different horizons is key to assist the management of call centers. In fact while forecasts at monthly horizon are needed for hiring new agents, forecasts at weekly and daily horizon are used for the scheduling of the available pool of agents (Akşin et al., 2007).

Focusing on individual forecasts, both univariate and multivariate symmetric losses ($\phi = 0.5$) suggest that the best performing model is the SARMAX–GARCH. This result holds also when including the combined forecasts in the ranking. As for the combining methods, ABMA based on the AIC seems to be the best available option. Notice that the best ABMA combinations are those based on sets of models that include the SARMAX forecasts with and without GARCH equation and the ARMAX–GARCH, namely those individual forecasts to which are associated some of the lowest individual and system losses.

When analyzing asymmetric losses, the ranking of models changes according to the incidence of over- and under-forecasting; nevertheless, we can confirm most of the results just highlighted for the symmetric case. Interestingly, in the multivariate case, when under-forecasting is more penalized than over-forecasting ($\phi = 0.58$), the MEM becomes the best option.

Table 5: Ranking of models and combined forecasts

	$h=1$			$h=7$			$h=28$			$h=1,...,28$		
	$\phi=0.42$	$\phi=0.5$	$\phi=0.58$	$\phi=0.42$	$\phi=0.5$	$\phi=0.58$	$\phi=0.42$	$\phi=0.5$	$\phi=0.58$	$\phi=0.42$	$\phi=0.5$	$\phi=0.58$
$\mathcal{M}_0$	44 (13)	32 (7)	43 (12)	44 (13)	32 (7)	43 (12)	44 (13)	32 (7)	43 (12)	29 (7)	43 (12)	43 (12)
$\mathcal{M}_1$	18 (4)	19 (4)	23 (4)	18 (4)	19 (4)	23 (4)	18 (4)	19 (4)	23 (4)	26 (5)	23 (5)	22 (6)
$\mathcal{M}_2$	21 (5)	14 (**3**)	15 (**3**)	21 (5)	14 (**3**)	15 (**3**)	21 (5)	14 (**3**)	15 (**3**)	15 (**3**)	14 (**3**)	28 (7)
$\mathcal{M}_3$	26 (8)	27 (6)	24 (5)	26 (8)	27 (6)	24 (5)	26 (8)	27 (6)	24 (5)	18 (4)	22 (4)	37 (10)
$\mathcal{M}_4$	13 (**3**)	5 (**2**)	4 (**2**)	13 (**3**)	5 (**2**)	4 (**2**)	13 (**3**)	5 (**2**)	4 (**2**)	8 (**2**)	6 (**2**)	31 (8)
$\mathcal{M}_5$	9 (**2**)	**1** (**1**)	**3** (**1**)	9 (**2**)	**1** (**1**)	**3** (**1**)	9 (**2**)	**1** (**1**)	**3** (**1**)	4 (**1**)	**1** (**1**)	34 (9)
$\mathcal{M}_6$	23 (6)	20 (5)	27 (6)	23 (6)	20 (5)	27 (6)	23 (6)	20 (5)	27 (6)	28 (6)	26 (6)	8 (**3**)
$\mathcal{M}_7$	5 (**1**)	43 (12)	45 (14)	5 (**1**)	43 (12)	45 (14)	5 (**1**)	43 (12)	45 (14)	44 (13)	45 (14)	45 (14)
$\mathcal{M}_8$	41 (12)	40 (11)	37 (9)	41 (12)	40 (11)	37 (9)	41 (12)	40 (11)	37 (9)	40 (11)	37 (9)	6 (**2**)
$\mathcal{M}_9$	37 (10)	36 (10)	33 (8)	37 (10)	36 (10)	33 (8)	37 (10)	36 (10)	33 (8)	37 (10)	34 (8)	14 (4)
$\mathcal{M}_{10}$	39 (11)	35 (9)	32 (7)	39 (11)	35 (9)	32 (7)	39 (11)	35 (9)	32 (7)	35 (9)	32 (7)	16 (5)
$\mathcal{M}_{11}$	45 (14)	44 (13)	40 (10)	45 (14)	44 (13)	40 (10)	45 (14)	44 (13)	40 (10)	43 (12)	42 (11)	**1** (**1**)
$\mathcal{M}_{12}$	29 (9)	45 (14)	44 (13)	29 (9)	45 (14)	44 (13)	29 (9)	45 (14)	44 (13)	45 (14)	44 (13)	44 (13)
$\mathcal{M}_{13}$	24 (7)	33 (8)	42 (11)	24 (7)	33 (8)	42 (11)	24 (7)	33 (8)	42 (11)	30 (8)	41 (10)	42 (11)
Avg. $\mathcal{G}_1$	16 [13]	21 [16]	12 [10]	16 [13]	21 [16]	12 [10]	16 [13]	21 [16]	12 [10]	25 [21]	18 [15]	9 [6]
Avg. $\mathcal{G}_2$	17 [14]	16 [13]	17 [14]	17 [14]	16 [13]	17 [14]	17 [14]	16 [13]	17 [14]	17 [14]	17 [14]	26 [20]
Avg. $\mathcal{G}_3$	6 [5]	12 [10]	11 [9]	6 [5]	12 [10]	11 [9]	6 [5]	12 [10]	11 [9]	12 [10]	11 [9]	19 [14]
Avg. $\mathcal{G}_4$	12 [10]	10 [8]	7 [5]	12 [10]	10 [8]	7 [5]	12 [10]	10 [8]	7 [5]	11 [9]	9 [7]	25 [19]
Avg. $\mathcal{G}_5$	40 [29]	39 [29]	36 [28]	40 [29]	39 [29]	36 [28]	40 [29]	39 [29]	36 [28]	39 [29]	36 [28]	11 [8]
Tr. Avg. $\mathcal{G}_1$	8 [7]	23 [18]	10 [8]	8 [7]	23 [18]	10 [8]	8 [7]	23 [18]	10 [8]	20 [16]	16 [13]	12 [9]
Tr. Avg. $\mathcal{G}_2$	25 [18]	22 [17]	19 [16]	25 [18]	22 [17]	19 [16]	25 [18]	22 [17]	19 [16]	21 [17]	21 [18]	27 [21]
Tr. Avg. $\mathcal{G}_3$	15 [12]	11 [9]	13 [11]	15 [12]	11 [9]	13 [11]	15 [12]	11 [9]	13 [11]	13 [11]	13 [11]	20 [15]
Tr. Avg. $\mathcal{G}_4$	19 [15]	9 [7]	5 [**3**]	19 [15]	9 [7]	5 [**3**]	19 [15]	9 [7]	5 [**3**]	10 [8]	8 [6]	24 [18]
Tr. Avg. $\mathcal{G}_5$	47 [33]	47 [33]	47 [33]	47 [33]	47 [33]	47 [33]	47 [33]	47 [33]	47 [33]	46 [32]	47 [33]	47 [33]

Notes: continued on next page

Table 5 (*continued*)
Ranking of models and combined forecasts

	$h=1$			$h=7$			$h=28$			$h=1,...,28$		
	$\phi=0.42$	$\phi=0.5$	$\phi=0.58$	$\phi=0.42$	$\phi=0.5$	$\phi=0.58$	$\phi=0.42$	$\phi=0.5$	$\phi=0.58$	$\phi=0.42$	$\phi=0.5$	$\phi=0.58$
Med. $\mathcal{G}_1$	4 [4]	17 [14]	18 [15]	4 [4]	17 [14]	18 [15]	4 [4]	17 [14]	18 [15]	24 [20]	19 [16]	10 [7]
Med. $\mathcal{G}_2$	20 [16]	18 [15]	19 [16]	20 [16]	18 [15]	19 [16]	20 [16]	18 [15]	19 [16]	19 [15]	20 [17]	23 [17]
Med. $\mathcal{G}_3$	7 [6]	8 [6]	14 [12]	7 [6]	8 [6]	14 [12]	7 [6]	8 [6]	14 [12]	14 [12]	12 [10]	18 [13]
Med. $\mathcal{G}_4$	14 [11]	4 [**3**]	6 [4]	14 [11]	4 [**3**]	6 [4]	14 [11]	4 [**3**]	6 [4]	9 [7]	7 [5]	21 [16]
Med. $\mathcal{G}_5$	36 [27]	36 [27]	33 [26]	36 [27]	36 [27]	33 [26]	36 [27]	36 [27]	33 [26]	36 [27]	33 [26]	13 [10]
Min $\mathcal{G}_1$	31 [22]	31 [25]	41 [31]	31 [22]	31 [25]	41 [31]	31 [22]	31 [25]	41 [31]	**1** [**1**]	40 [31]	46 [32]
Min $\mathcal{G}_2$	**3** [**3**]	13 [11]	22 [19]	**3** [**3**]	13 [11]	22 [19]	**3** [**3**]	13 [11]	22 [19]	7 [6]	10 [8]	38 [28]
Min $\mathcal{G}_3$	**1** [**1**]	6 [4]	9 [7]	**1** [**1**]	6 [4]	9 [7]	**1** [**1**]	6 [4]	9 [7]	**2** [**2**]	4 [**3**]	40 [30]
Min $\mathcal{G}_4$	**2** [**2**]	7 [5]	8 [6]	**2** [**2**]	7 [5]	8 [6]	**2** [**2**]	7 [5]	8 [6]	**3** [**3**]	5 [4]	39 [29]
Min $\mathcal{G}_5$	35 [26]	34 [26]	30 [24]	35 [26]	34 [26]	30 [24]	35 [26]	34 [26]	30 [24]	34 [26]	31 [25]	17 [12]
Max $\mathcal{G}_1$	46 [32]	46 [32]	46 [32]	46 [32]	46 [32]	46 [32]	46 [32]	46 [32]	46 [32]	47 [33]	46 [32]	41 [31]
Max $\mathcal{G}_2$	32 [23]	28 [22]	28 [22]	32 [23]	28 [22]	28 [22]	32 [23]	28 [22]	28 [22]	31 [23]	28 [22]	4 [**3**]
Max $\mathcal{G}_3$	34 [25]	30 [24]	31 [25]	34 [25]	30 [24]	31 [25]	34 [25]	30 [24]	31 [25]	33 [25]	30 [24]	**2** [**1**]
Max $\mathcal{G}_4$	33 [24]	29 [23]	29 [23]	33 [24]	29 [23]	29 [23]	33 [24]	29 [23]	29 [23]	32 [24]	29 [23]	**3** [**2**]
Max $\mathcal{G}_5$	43 [31]	42 [31]	39 [30]	43 [31]	42 [31]	39 [30]	43 [31]	42 [31]	39 [30]	42 [31]	39 [30]	5 [4]
SIC $\mathcal{G}_2$	28 [20]	26 [21]	26 [21]	28 [20]	26 [21]	26 [21]	28 [20]	26 [21]	26 [21]	23 [19]	25 [20]	36 [27]
SIC $\mathcal{G}_3$	30 [21]	24 [19]	21 [18]	30 [21]	24 [19]	21 [18]	30 [21]	24 [19]	21 [18]	27 [22]	27 [21]	30 [23]
SIC $\mathcal{G}_4$	27 [19]	25 [20]	25 [20]	27 [19]	25 [20]	25 [20]	27 [19]	25 [20]	25 [20]	22 [18]	24 [19]	35 [26]
SIC $\mathcal{G}_5$	41 [30]	40 [30]	37 [29]	41 [30]	40 [30]	37 [29]	41 [30]	40 [30]	37 [29]	40 [30]	37 [29]	6 [5]
AIC $\mathcal{G}_2$	21 [17]	14 [12]	15 [13]	21 [17]	14 [12]	15 [13]	21 [17]	14 [12]	15 [13]	15 [13]	14 [12]	28 [22]
AIC $\mathcal{G}_3$	10 [8]	**1** [**1**]	**1** [**1**]	10 [8]	**1** [**1**]	**1** [**1**]	10 [8]	**1** [**1**]	**1** [**1**]	5 [4]	**2** [**1**]	32 [24]
AIC $\mathcal{G}_4$	10 [8]	**1** [**1**]	**1** [**1**]	10 [8]	**1** [**1**]	**1** [**1**]	10 [8]	**1** [**1**]	**1** [**1**]	5 [4]	**2** [**1**]	32 [24]
AIC $\mathcal{G}_5$	37 [28]	36 [27]	33 [26]	37 [28]	36 [27]	33 [26]	37 [28]	36 [27]	33 [26]	37 [28]	34 [27]	14 [11]

Notes: the first column uses the following shorthand notation: "Avg." = Average, "Tr. Avg." = Trimmed Average, "Med." = Median, "SIC" = ABMA using SIC and "AIC" = ABMA using AIC; entries represent the ranking of models and forecast combinations based on the statistics $\sqrt{P^{-1}\sum_{t=1}^{P}\mathcal{L}_t}$, where $\mathcal{L}$ is the generalized loss function in Eq. (2) and $P=351$; statistics in columns 2-10 are based on the univariate loss for forecast horizons, $h=1,7,28$, while those in the last three columns are based on the multivariate loss for $h=1,...,28$; the shape of the loss function is determined by $\rho=2$ and the asymmetry coefficient: $\phi=0.42,0.50,0.58$ ($\tau=0.16,0.00,-0.16$); these values guarantee that the multivariate loss is always non-negative (see Komunjer and Owyang (2012) for details); entries outside brackets represent the overall ranking of the forecast; entries in round brackets represent the ranking among models, while entries in square brackets denote the ranking among combining methods; models, $\mathcal{M}_m$, groups of models, $\mathcal{G}_i$, and combining methods are described in Tables 1, 2 and 3; entries in bold denote the three models with the lowest loss.

Table 6: Reality Check test

	$h=1$			$h=7$			$h=28$		
	$\phi=0.42$	$\phi=0.5$	$\phi=0.58$	$\phi=0.42$	$\phi=0.5$	$\phi=0.58$	$\phi=0.42$	$\phi=0.5$	$\phi=0.58$
$\mathcal{M}_0$	–	–	–	⋆	–	–	⋆⋆	⋆⋆	⋆⋆
$\mathcal{M}_1$	–	–	⋆	–	–	–	–	–	–
$\mathcal{M}_2$	–	–	⋆	⋆	⋆	⋆	–	–	⋆
$\mathcal{M}_3$	–	–	⋆	–	–	–	–	⋆	⋆
$\mathcal{M}_4$	–	–	⋆⋆	⋆⋆	⋆⋆	⋆⋆	–	⋆	⋆⋆
$\mathcal{M}_5$	–	⋆⋆	⋆⋆	⋆⋆	⋆⋆	⋆⋆	⋆⋆	⋆⋆	⋆⋆
$\mathcal{M}_6$	⋆⋆	⋆⋆	⋆⋆	⋆	⋆⋆	⋆⋆	–	–	⋆
$\mathcal{M}_7$	⋆	⋆⋆	⋆⋆	–	–	–	⋆	⋆	⋆
$\mathcal{M}_8$	–	–	–	–	–	–	–	–	–
$\mathcal{M}_9$	–	–	–	–	–	–	–	–	–
$\mathcal{M}_{10}$	–	–	–	–	–	–	–	–	–
$\mathcal{M}_{11}$	–	–	–	–	–	–	–	–	–
$\mathcal{M}_{12}$	–	–	–	–	–	–	–	–	–
$\mathcal{M}_{13}$	–	–	–	–	–	–	⋆⋆	⋆⋆	⋆⋆
Avg. $\mathcal{G}_1$	–	–	⋆	–	⋆	⋆⋆	–	⋆	⋆⋆
Avg. $\mathcal{G}_2$	–	⋆⋆	⋆⋆	–	–	–	–	–	–
Avg. $\mathcal{G}_3$	⋆⋆	⋆⋆	⋆⋆	–	–	⋆	–	–	–
Avg. $\mathcal{G}_4$	–	⋆⋆	⋆⋆	⋆	⋆	⋆⋆	⋆⋆	⋆⋆	⋆⋆
Avg. $\mathcal{G}_5$	–	–	–	–	–	–	–	–	⋆
Tr. Avg. $\mathcal{G}_1$	–	⋆	⋆	⋆	–	⋆	–	⋆	⋆⋆
Tr. Avg. $\mathcal{G}_2$	–	⋆	⋆⋆	–	⋆	⋆	–	–	–
Tr. Avg. $\mathcal{G}_3$	–	⋆	⋆⋆	⋆	⋆	⋆⋆	–	–	⋆⋆
Tr. Avg. $\mathcal{G}_4$	–	–	–	⋆	–	–	⋆⋆	⋆⋆	⋆⋆
Tr. Avg. $\mathcal{G}_5$	–	–	–	–	–	–	–	–	–
Med. $\mathcal{G}_1$	⋆	⋆⋆	⋆⋆	–	⋆⋆	⋆⋆	–	–	–
Med. $\mathcal{G}_2$	–	–	–	–	–	⋆	–	–	–
Med. $\mathcal{G}_3$	–	⋆⋆	⋆⋆	⋆⋆	⋆⋆	⋆⋆	–	–	⋆⋆
Med. $\mathcal{G}_4$	–	⋆	⋆⋆	⋆⋆	⋆⋆	⋆⋆	⋆⋆	⋆⋆	⋆⋆
Med. $\mathcal{G}_5$	–	–	–	–	–	–	–	–	–
Min $\mathcal{G}_1$	⋆⋆	⋆⋆	⋆⋆	–	–	–	⋆⋆	⋆⋆	⋆⋆
Min $\mathcal{G}_2$	⋆⋆	⋆⋆	⋆⋆	⋆	⋆	⋆⋆	⋆⋆	⋆⋆	⋆⋆
Min $\mathcal{G}_3$	⋆⋆	⋆⋆	⋆⋆	⋆⋆	⋆⋆	⋆⋆	⋆⋆	⋆⋆	⋆⋆
Min $\mathcal{G}_4$	⋆⋆	⋆⋆	⋆⋆	⋆⋆	⋆⋆	⋆⋆	⋆⋆	⋆⋆	⋆⋆
Min $\mathcal{G}_5$	–	–	–	–	–	–	–	–	–
Max $\mathcal{G}_1$	–	–	–	–	–	–	–	–	–
Max $\mathcal{G}_2$	–	–	–	–	–	–	–	–	⋆
Max $\mathcal{G}_3$	–	–	–	–	–	–	–	–	–
Max $\mathcal{G}_4$	–	–	–	–	–	–	–	–	–
Max $\mathcal{G}_5$	–	–	–	–	–	–	–	–	–
SIC $\mathcal{G}_2$	–	–	–	–	⋆	⋆	–	–	⋆
SIC $\mathcal{G}_3$	–	–	–	–	–	–	–	⋆	⋆⋆
SIC $\mathcal{G}_4$	–	–	–	–	–	⋆	–	–	⋆
SIC $\mathcal{G}_5$	–	–	–	–	–	–	–	–	–
AIC $\mathcal{G}_2$	–	–	⋆	⋆	⋆⋆	⋆⋆	–	⋆	⋆
AIC $\mathcal{G}_3$	–	⋆⋆	⋆⋆	⋆⋆	⋆⋆	⋆⋆	⋆⋆	⋆⋆	⋆⋆
AIC $\mathcal{G}_4$	–	⋆⋆	⋆⋆	⋆⋆	⋆⋆	⋆⋆	⋆⋆	⋆⋆	⋆⋆
AIC $\mathcal{G}_5$	–	–	–	–	–	–	–	–	–

Notes: the table presents results of the Reality Check test of White (2000) as modified by Hansen (2005); the benchmark model is indicated in the first column, where the following shorthand notation is used: "Avg." = Average, "Tr. Avg." = Trimmed Average, "Med." = Median, "SIC" = ABMA using SIC and "AIC" = ABMA using AIC; the test is implemented using the stationary (block) bootstrap of Politis and Romano (1994); the number of bootstrap repetitions is equal to 999, the block length equals 29 days; a p-value lower than 0.05 indicates that we reject the hypothesis that the benchmark performs as well as the best alternative model; "–" denotes a p-value < 0.05, "⋆" denotes $0.05 \leq$ p-value < 0.1, "⋆⋆" denotes a p-value ≥ 0.1.

Focusing on combination schemes, "minimum forecasts" based on $\mathcal{G}_3$ and $\mathcal{G}_4$ yield the lowest average losses when $\phi = 0.42$. On the contrary, when under-forecasting in costlier than over-forecasting ($\phi = 0.52$), these forecasts do not make in the first positions anymore. In this case, either the ABMA-AIC combining methods, or the "maximum forecasts" lead to the lowest average losses. Moreover, we can observe that neither count data models, nor the Spline-SARX model seem to be valuable options.

Table 7: In Model Confidence Set?

	$h=1$			$h=7$			$h=28$		
	$\phi=0.42$	$\phi=0.5$	$\phi=0.58$	$\phi=0.42$	$\phi=0.5$	$\phi=0.58$	$\phi=0.42$	$\phi=0.5$	$\phi=0.58$
$\mathcal{M}_0$	–	–	–	–	–	–	–	–	–
$\mathcal{M}_1$	–	–	–	–	–	–	–	–	–
$\mathcal{M}_2$	–	–	–	–	–	–	–	–	–
$\mathcal{M}_3$	–	–	–	–	–	–	–	–	–
$\mathcal{M}_4$	–	✓	–	–	–	–	–	–	–
$\mathcal{M}_5$	–	✓	✓	–	–	–	–	–	–
$\mathcal{M}_6$	–	✓	✓	–	–	–	–	–	–
$\mathcal{M}_7$	✓	✓	✓	–	–	–	–	–	–
$\mathcal{M}_8$	–	–	–	–	–	–	–	–	–
$\mathcal{M}_9$	–	–	–	–	–	–	–	–	–
$\mathcal{M}_{10}$	–	–	–	–	–	–	–	–	–
$\mathcal{M}_{11}$	–	–	–	–	–	–	–	–	–
$\mathcal{M}_{12}$	–	–	–	–	–	–	–	–	–
$\mathcal{M}_{13}$	–	–	–	–	–	–	–	–	–
Avg. $\mathcal{G}_1$	–	–	✓	–	–	–	–	–	–
Avg. $\mathcal{G}_2$	–	✓	✓	–	–	–	–	–	–
Avg. $\mathcal{G}_3$	–	✓	✓	–	–	–	–	–	–
Avg. $\mathcal{G}_4$	–	✓	✓	–	–	–	–	–	–
Avg. $\mathcal{G}_5$	–	–	–	–	–	–	–	–	–
Tr. Avg. $\mathcal{G}_1$	–	✓	✓	–	–	–	–	–	–
Tr. Avg. $\mathcal{G}_2$	–	–	–	–	–	–	–	–	–
Tr. Avg. $\mathcal{G}_3$	–	–	–	–	–	–	–	–	–
Tr. Avg. $\mathcal{G}_4$	–	–	–	–	–	–	–	–	–
Tr. Avg. $\mathcal{G}_5$	–	–	–	–	–	–	–	–	–
Med. $\mathcal{G}_1$	–	✓	✓	–	–	–	–	–	–
Med. $\mathcal{G}_2$	–	–	–	–	–	–	–	–	–
Med. $\mathcal{G}_3$	–	✓	✓	–	–	–	–	–	–
Med. $\mathcal{G}_4$	–	–	–	–	–	–	–	–	–
Med. $\mathcal{G}_5$	–	–	–	–	–	–	–	–	–
Min $\mathcal{G}_1$	–	–	–	–	–	–	–	–	–
Min $\mathcal{G}_2$	✓	✓	✓	–	–	–	–	–	–
Min $\mathcal{G}_3$	✓	✓	✓	✓	–	–	✓	–	–
Min $\mathcal{G}_4$	✓	✓	✓	✓	–	–	✓	–	–
Min $\mathcal{G}_5$	–	–	–	–	–	–	–	–	–
Max $\mathcal{G}_1$	–	–	–	–	–	–	–	–	–
Max $\mathcal{G}_2$	–	–	–	–	–	–	–	–	–
Max $\mathcal{G}_3$	–	–	–	–	–	–	–	–	–
Max $\mathcal{G}_4$	–	–	–	–	–	–	–	–	–
Max $\mathcal{G}_5$	–	–	–	–	–	–	–	–	–
SIC $\mathcal{G}_2$	–	–	–	–	–	–	–	–	–
SIC $\mathcal{G}_3$	–	–	–	–	–	–	–	–	–
SIC $\mathcal{G}_4$	–	–	–	–	–	–	–	–	–
SIC $\mathcal{G}_5$	–	–	–	–	–	–	–	–	–
AIC $\mathcal{G}_2$	–	–	–	–	–	–	–	–	–
AIC $\mathcal{G}_3$	–	–	–	–	–	–	–	–	–
AIC $\mathcal{G}_4$	–	✓	✓	✓	✓	✓	✓	✓	✓
AIC $\mathcal{G}_5$	–	–	–	–	–	–	–	–	–

Notes: the table presents results of the Model Confidence Set (MCS) procedure of Hansen et al. (2011) implemented using the stationary (block) bootstrap of Politis and Romano (1994); the number of bootstrap repetitions is equal to 999, the block length equals 29 days; "–" indicates that the model is not in the MCS at the 90% confidence level, while "✓" indicates that the model belongs to the MCS; the first column uses the following shorthand notation: "Avg." = Average, "Tr. Avg." = Trimmed Average, "Med." = Median, "SIC" = ABMA using SIC and "AIC" = ABMA using AIC.

Lastly, we find that second–moment modeling is important when forecasting call arrivals. In fact, when a GARCH component is added to ARMAX and SARMAX models their ranking improves in most cases. This result suggest that these simple models could also been useful in the literature dealing with density forecasts (see e.g. Taylor, 2012).

From the standing point of a practitioner, results in Table 5 also suggest that, independently of the shape of the loss function, outsourcing the forecasting exercise could be worth

its cost; in fact, the benchmark SRW model is always outperformed by other relatively more sophisticated specifications.[5]

Since the number of forecasts under consideration is quite large, these conclusions might, to some extent, be subject to data snooping effects. Both the Reality Check test (RCT) of White (2000) and the Model Confidence Set (MCS) of Hansen et al. (2011), are designed to deal with data snooping. The difference between the two procedures is that while the former requires a benchmark model, the latter does not.

Results of the RCT are shown in Table 6; the null hypothesis is that the benchmark performs as well as the best alternative model. The results are based on the consistent p-values of Hansen (2005), who has shown that the original procedure has low power when a poor performing forecast enters the set of alternative models.

Using the SARMAX–GARCH model, or the ABMA–AIC combined forecasts (based either on $\mathcal{G}_3$, or on $\mathcal{G}_4$), we reject the null hypothesis only once.

The MCS is used to compare the forecast accuracy of models without selecting a benchmark model and yields a set of specifications that contains the best forecast with a prespecified asymptotic probability. As it can be seen from Table 7, this test is more selective than the RCT. Considering one day ahead forecasts and under MSE loss the MCS, at the 90% confidence level, contains only four individual models: SARMAX, SARMAX–GARCH, PAR and the Airline model. When over-forecasting is more penalized than under-prediction (i.e. $\phi = 0.42$), the only model entering the MCS is the Airline; while, when positive forecast errors are more heavily weighted than negative forecast errors (i.e. $\phi = 0.58$), the SARMAX–GARCH and the PAR are also in the MCS. When the forecast horizon is one week, or one month, only the ABMA–AIC combined forecasts based on $\mathcal{G}_4$ are always in the MCS.

Therefore, when the loss function is parametrized so as to penalize under-prediction more heavily than over-prediction, the best performing model and combination method are the SARMAX–GARCH and the ABMA–AIC based on $\mathcal{G}_4$.

[5]This fact is supported also by the Diebold and Mariano (1995) test. These results are available from the authors upon request.

3.2 Choosing the best forecasting method

We now focus on the task of selecting the best method given a forecast horizon of one day and assuming that the manager of the call center is more adverse to under-staffing than to over-staffing. We thus shrink the set of alternative forecasts so as to include all individual models and the combined forecast obtained with ABMA-AIC applied to group $\mathcal{G}_4$.

The economic evaluation of forecasts based on the willingness to pay, δ_i, and the incremental value of information, ΔV_i, is presented in Table 8. Although both economic measures of performance decrease as the absolute risk aversion increases, the manager's willingness to pay seems to be less responsive to such a change than the incremental value of information.

The second column of Table 8 shows the percentage change in the root MSE distance for comparison: an entry below 100 indicates that the i-th model outperforms the benchmark. All measures suggest that the worst performing model is the MEM; as for the best model, both money metrics point to the SARMAX–GARCH and to the ABMA–AIC combined forecast. The manager is willing to pay up to 1687 Euro in order to use these models instead of the benchmark. The model to which is associated the minimum (positive) willingness to pay, 912 Euro, is the Poisson count data specification.

The ranking based on the incremental value of information is consistent with that based on the willingness to pay. On the contrary, being symmetric about zero forecast errors, the ranking based on the MSE ranking is quite different: for instance, the Airline model would be preferred to the SARMAX–GARCH model which is the best option when the loss function is consistent with the manager's compensation scheme.

Overall, both economic and statistical evaluation of models indicate the SARMAX–GARCH model and the ABMA-AIC combining method based on $\mathcal{G}_4$ as the best options for the manager. However, using only statistical methods, we cannot clearly identify which of these options is the best. On the contrary, given that the monetary value of the two forecasts and the manager's willingness to pay for them are very similar, we can conclude that the SARMAX–GARCH is to be preferred to the ABMA–AIC combined method.

Actually, given that the latter method involves more than one model and that the specification of models has to be periodically revised, it will have higher maintenance costs

Table 8: Economic evaluation of models

		δ_i (Euro)			ΔV_i (Euro)		
	ΔRMSE (%)	$\lambda = 0.0002$	$\lambda = 0.0003$	$\lambda = 0.0005$	$\lambda = 0.0002$	$\lambda = 0.0003$	$\lambda = 0.0005$
$\mathcal{M}_1$	58.40	1377.40	1377.40	1377.30	401.66	114.03	11.96
$\mathcal{M}_2$	58.50	1544.90	1544.80	1544.60	445.01	125.19	12.97
$\mathcal{M}_3$	62.27	1342.40	1342.40	1342.30	392.47	111.63	11.74
$\mathcal{M}_4$	57.06	1413.70	1413.70	1413.70	411.15	116.49	12.19
$\mathcal{M}_5$	56.85	1687.40	1687.30	1687.20	481.04	134.29	13.77
$\mathcal{M}_6$	59.02	1433.70	1433.60	1433.50	416.35	117.84	12.31
$\mathcal{M}_7$	55.86	1517.40	1517.40	1517.30	437.98	123.40	12.81
$\mathcal{M}_8$	76.36	912.39	912.30	912.16	275.33	80.22	8.73
$\mathcal{M}_9$	76.10	988.59	988.47	988.28	296.65	86.05	9.30
$\mathcal{M}_{10}$	76.10	988.59	988.47	988.28	296.65	86.05	9.30
$\mathcal{M}_{11}$	123.66	-933.48	-933.27	-932.94	-324.07	-105.60	-13.59
$\mathcal{M}_{12}$	62.72	1031.10	1031.10	1030.90	308.44	89.26	9.62
$\mathcal{M}_{13}$	61.19	1217.50	1217.40	1217.40	359.21	102.88	10.93
AIC $\mathcal{G}_4$	56.85	1687.40	1687.30	1687.20	481.04	134.29	13.77

Notes: economic evaluation of one day ahead forecasts; ΔRMSE=100$\times$(RMSE$_i$/RMSE$_{\mathcal{M}_0}$), where the RMSE corresponds to the flexible loss distance for $\rho = 2$ and $\phi = 0.5$; the incremental value of information is $\Delta V_i = V_i - V_{\mathcal{M}_0}$, where V_i is the value of information from model i; the willingness to pay for model i is $\delta = CE_i - CE_{\mathcal{M}_0}$, where CE_i is the certainty equivalent from model i; λ is the coefficient of risk aversion.

than the SARMAX–GARCH model. Moreover, if the model is run by an employee of the call center and not by the professional forecaster, we see the "ease-of-use" as a critical factor for the choice of the best forecast.

All in all, we have shown that simple measures of performance expressed in monetary terms are easy to construct and offer greater flexibility than the often used symmetric loss functions. This flexibility allows the forecasts' user and the adviser to judge the predictive performance of models with the same metric. Moreover, being expressed in monetary terms, we believe that these measures are more interesting for practitioners than traditional statistical distances. Finally, from the perspective of the professional forecaster, we see the results in this section as complementary to those based on flexible loss functions. From the standpoint of forecast's users, we believe that the economic measures are to be preferred because more closely linked to their profit maximizing behavior.

3.3 Robustness checks: alternative dataset

The usefulness of our results clearly depends on their degree of generalizabililty. To shed light on this matter we consider two alternative series representing the number of call arrivals recorded at the call centers operated by two banks, one located in Israel and the other in the U.S.. In Table 9 we list for each series the model with the lowest Mean Squared Forecast Error. As it can been seen either the ARMAX–GARCH or SARMAX–GARCH are the winning option in all countries and at all forecast horizons.

Table 9: MSE ranking of models in Israel, Italy and the U.S..

country	$h = 1$	$h = 7$	$h = 21$	$h = 1, ..., 28$
Israel	$\mathcal{M}_2$	$\mathcal{M}_5$	$\mathcal{M}_2$	$\mathcal{M}_2$
Italy	$\mathcal{M}_5$	$\mathcal{M}_5$	$\mathcal{M}_5$	$\mathcal{M}_5$
U.S.	$\mathcal{M}_2$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{M}_2$

Notes: the first column identifies the dataset used; entries in columns 2-4 represent the model with the lowest Mean Squared Forecast Error (MSFE) at forecast horizon 1,7,28 days, while entries in the last column represent the model with the lowest sum of MSFE over forecast horizons $h = 1, 2, \ldots, 28$. Models are described in Tables 1.

4 Conclusions

Call centers' managers and companies relying on call center services are interested in obtaining accurate forecasts of call arrivals for achieving optimal operating efficiency. This paper has shown how to choose among forecasting methods in call centers.

The empirical exercise in this paper mimics the interaction between a professional forecaster and a manager needing forecasts of incoming calls to decide how many agents are required each day at a call center. In this context, we have evaluated fourteen models and a set of seven forecast combination schemes using flexible loss functions, statistical tests and economic measures of performance.

Each of these forecasts is able to capture one or more key features of the daily call arrival series. Moreover, all of the models and combination methods are computationally tractable, with a relatively small number of parameters that can be easily estimated and updated with any off-the-shelf statistical software as new data become available. This is a crucial point in the selection of a model for forecasting call arrivals. In fact, to be of practical use, it must not only reproduce the key features of the data, but also be easily implementable so as to quickly generate new forecasts to update the operational decisions in call centers (Ibrahim et al., 2016).

After taking a stance on the shape of the statistical loss function, parametrized so as to make under-forecasting costlier than over-forecasting and to be consistent with the compensation scheme used to pay the manager of the call center, we have shown that the professional forecaster can shrink the number of methods to present to the customer by using the Reality Check test and the Model Confidence Set.

These tests as well as the ranking of models suggest that the best available options are the SARMAX–GARCH and a combined forecast obtained with ABMA. Subsequently, we have

shown that the economic evaluation of forecast accuracy leads to the same results. However, given that individual and combined forecasts have approximately the same monetary value, the manager will choose the SARMAX–GARCH model, due to lower maintenance costs. The maintenance costs of a forecasting model used by an employee of the call center include direct costs, due to periodical checks of the specification, as well as indirect costs, associated to the relative complexity of the forecasting method. Given that a combined forecast involves a set of models, whose specifications have to be periodically checked, it will probably lead to higher maintenance costs and hence the SARMAX–GARCH model will be the best available choice.

We have presented a wide array of results involving different loss functions, forecast horizons and call arrival series, we can also draw some more general conclusions. First, it emerges that from the point of view of a manager, outsourcing the forecasting exercise could be worth its cost; in fact, the benchmark SRW model is always outperformed by other relatively more sophisticated specifications. Second, independently of the forecast horizon and for any shape of the loss function, the combination of forecasts, especially if based on optimal combining weights calculated by means of ABMA, proves useful and lead to lower statistical losses than most individual models. Third, the statistical evaluation of models indicates that second–moment modeling, and not only seasonality, is important; in fact either the ARMAX–GARCH or SARMAX–GARCH model emerges as one of the best alternatives both among individual and combined forecasts. This last result implies that anticipating the variability of call arrivals is extremely important, in that when a call centers operates under a SLA, higher uncertainty requires higher staffing levels to meet service quality objectives. Moreover, given its ease-of-implementation the SARMAX–GRACH model seems to be a good candidate also for density forecasting.

References

Akşin, Z., Armony, M., and Mehrotra, V. (2007). The modern call center: A multi-disciplinary perspective on operations management research. *Production and Operations Management*, 16:665–688.

Bastianin, A., Galeotti, M., and Manera, M. (2011). Forecast evaluation in call centers: combined forecasts, flexible loss functions and economic criteria. Departmental Working Papers 2011-08, Department of Economics, Management and Quantitative Methods at Universit degli Studi di Milano.

Collender, R. N. and Chalfant, J. A. (1986). An alternative approach to decisions under uncertainty using the empirical moment-generating function. *American Journal of Agricultural Economics*, 68(3):727–731.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–63.

Dorfman, J. H. and McIntosh, C. S. (1997). Economic criteria for evaluating commodity price forecasts. *Journal of Agricultural and Applied Economics*, 29(2):337–345.

Elbasha, E. H. (2005). Risk aversion and uncertainty in cost-effectiveness analysis: the expected-utility, moment-generating function approach. *Health Economics*, 14(5):457–470.

Elliott, G., Komunjer, I., and Timmermann, A. (2005). Estimation and testing of forecast rationality under flexible loss. *Review of Economic Studies*, 72(4):1107–1125.

Engle, R. F. (2002). New frontiers for ARCH models. *Journal of Applied Econometrics*, 17(5):425–446.

Franses, P. H. and van Dijk, D. (2005). The forecasting performance of various models for seasonality and nonlinearity for quarterly industrial production. *International Journal of Forecasting*, 21(1):87–102.

Gans, N., Koole, G., and Mandelbaum, A. (2003). Telephone call centers: A tutorial and literature review. *Manufacturing and Service Operations Management*, 5(2):79–141.

Gardner, Jr., E. S. (2006). Exponential smoothing: The state of the art–part II. *International Journal of Forecasting*, 22(4):637–666.

Garratt, A., Lee, K., Pesaran, H. M., and Shin, Y. (2003). Forecast uncertainties in macroeconomic modeling: An application to the U.K. economy. *Journal of the American Statistical Association*, 98(464):829–838.

Gbur, E. E. and Collins, R. A. (1989). A small-sample comparison of estimators in the EU-MGF approach to decision making. *American Journal of Agricultural Economics*, 71(1):202–210.

Hansen, P. R. (2005). A test for superior predictive ability. *Journal of Business & Economic Statistics*, 23(4):365–380.

Hansen, P. R., Lunde, A., and Nason, J. M. (2011). The model confidence set. *Econometrica*, 79(2):453–497.

Ibrahim, R., Ye, H., Ł'Ecuyer, P., and Shen, H. (2016). Modeling and forecasting call center arrivals: A literature survey and a case study. *International Journal of Forecasting*, 32(3):865–874.

Jung, R. C. and Tremayne, A. R. (2011). Useful models for time series of counts or simply wrong ones? *AStA Advances in Statistical Analysis*, 95(1):59–91.

Komunjer, I. and Owyang, M. T. (2012). Multivariate forecast evaluation and rationality testing. *Review of Economics and Statistics*, 94(4):1066–1080.

Leitch, G. and Tanner, J. E. (1991). Economic forecast evaluation: Profits versus the conventional error measures. *The American Economic Review*, 81(3):580–590.

Politis, D. N. and Romano, J. P. (1994). The stationary bootstrap. *Journal of the American Statistical Association*, 89(428):1303–1313.

Shen, H. and Huang, J. Z. (2008). Forecasting time series of inhomogeneous Poisson processes with application to call center workforce management. *Annals of Applied Statistics*, 2(2):601–623.

Stock, J. H. and Watson, M. W. (2004). Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting*, 23(6):405–430.

Stolletz, R. (2003). *Performance Analysis and Optimization of Inbound Call Centers*. Springer Berlin Heidelberg.

Taylor, J. W. (2008). A comparison of univariate time series methods for forecasting intraday arrivals at a call center. *Management Science*, 54(2):253–265.

Taylor, J. W. (2012). Density forecasting of intraday call center arrivals using models based on exponential smoothing. *Management Science*, 58(3):534–549.

Timmermann, A. (2006). Forecast combinations. In Elliott, G., Granger, C. W. J., and Timmermann, A., editors, *Handbook of Economic Forecasting*, volume 1, chapter 4, pages 135–196. Elsevier, Amsterdam.

White, H. (2000). A reality check for data snooping. *Econometrica*, 68(5):1097–1126.

Zeng, T. and Swanson, N. R. (1998). Predictive evaluation of econometric forecasting models in commodity futures markets. *Studies in Nonlinear Dynamics & Econometrics*, 2(4):159–177.