forecasting

MDPI

Article

Forecasting Demand Under Limited Data: Benchmark Models for Workforce Capacity Planning in Translation Services

Lorena Hernández-Mastrapa [1,2], Daniel René Tasé-Velázquez [2,3], Renato Máximo-Sátiro [4], Gelmar García-Vidal [5], Alexander Sánchez-Rodríguez [6,*] and Reyner Pérez-Campdesuñer [2,5]

1 Graduate Program in Production Engineering, São Carlos School of Engineering, University of São Paulo, São Paulo 05508-220, Brazil; lorenahmastrapa@gmail.com
2 General Directorate of Postgraduate Studies, Universidad UTE, Quito 170508, Ecuador; danielvelazquez@fho.edu.br (D.R.T.-V.); reyner.perez@ute.edu.ec (R.P.-C.)
3 Hermínio Ometto Foundation University Center—FHO, São Paulo 13607-339, Brazil
4 Accounting and Economics (FACE), Faculty of Administration, Federal University of Goiás (UFG), Goiânia 74690-900, Brazil; r.maximo.satiro@gmail.com
5 Faculty of Law, Administrative and Social Sciences, Universidad UTE, Quito 170527, Ecuador; gelmar.garcia@ute.edu.ec
6 Faculty of Engineering Sciences and Industries, Universidad UTE, Quito 170527, Ecuador
* Correspondence: alexander.sanchez@ute.edu.ec; Tel.: +593-987171408

Highlights

What are the main findings?

- The Mean benchmark produced the lowest MAE, RMSE, MASE, and sMAPE across one-, five-, and twenty-working-day forecast horizons.
- The final Mean forecast of 5472.12 words per working day implied a baseline requirement of three translators under the nominal productivity assumption, although observed demand exceeded three-translator capacity on 34.43% of working days.

What are the implications of the main findings?

- Forecasting models for workforce planning should be selected according to the operational consequences of their errors rather than a single accuracy measure.
- Parsimonious benchmarks can support capacity planning when historical data are limited, but contingency resources remain necessary to manage short-term demand peaks.

Abstract

Demand variability complicates workforce capacity planning in knowledge-intensive services, particularly when only short historical records are available. This study evaluates parsimonious benchmark models for forecasting translation-service demand and translating the resulting forecasts into staffing requirements. The empirical analysis used 122 daily observations collected over six months from a translation team providing services in English and Spanish. Demand was measured as the number of words requested per working day, while nominal individual capacity was operationalized as 2000 translated words per day. Three benchmark forecasting methods—Naive, Mean, and Drift—were compared through expanding-window rolling-origin evaluation using ME, MAE, RMSE, MASE, and sMAPE. At the one-working-day horizon, the Mean benchmark achieved the lowest MAE (3221.66 words), RMSE (4424.81 words), MASE (0.845), and sMAPE (58.14%), and it maintained the lowest values for these measures at five- and twenty-working-day horizons. Re-estimated using all 122 observations, the Mean benchmark generated a point forecast of 5472.12 words per working day, equivalent to 2.74 translator-equivalents and a baseline

Check for updates

Academic Editors: Foteini Kyriazi, Dimitrios D. Thomakos and Stelios Arvanitis

Received: 2 August 2026
Revised: 9 September 2026
Accepted: 10 September 2026
Published: 13 September 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

requirement of three translators under the nominal productivity assumption. However, observed demand exceeded three-translator capacity on 34.43% of working days, indicating the need for flexible contingency capacity. The study provides a transparent framework connecting benchmark forecast evaluation with workforce-capacity decisions under limited temporal coverage.

Keywords: parsimonious forecasting; short time series; forecast accuracy; service operations; workforce planning; capacity–demand alignment; knowledge-intensive services

1. Introduction

Balancing demand and capacity is a persistent challenge in service operations. Unlike manufacturing environments, in which inventories can partially absorb fluctuations between supply and demand, many services cannot store productive capacity for later use [1–3]. Capacity shortages may therefore generate delays, excessive workloads, and service deterioration, whereas excess capacity produces idle time and underutilized resources. These tensions are particularly important in knowledge-intensive services, where capacity is embedded in specialized personnel and cannot always be expanded rapidly through recruitment, training, or reassignment [4–6]. Operations-management research further shows that variability becomes increasingly disruptive as utilization approaches effective capacity, amplifying queues, delays, and accumulated work [7], Chs. 8–9; [8], Ch. 11.

Translation services provide a clear example of this problem. Incoming requests vary in volume, timing, complexity, language direction, and urgency, while productive capacity depends primarily on the availability and effective output of skilled translators. Short-term workload increases must therefore be absorbed by the available team, postponed, or assigned to external providers, whereas permanent staffing based on occasional peaks may generate substantial underutilization during lower-demand periods. Reliable demand information is consequently important for workforce sizing, task allocation, delivery commitments, and contingency-capacity decisions.

Operational data can support these decisions only when they are consistently recorded, transformed into comparable measures, and analyzed in a form suitable for managerial use [9–11]. Forecasting provides the connection between historical demand and forward-looking resource decisions by estimating future workloads and their uncertainty [12,13]. In the present study, the analytical purpose is predictive rather than explanatory: the objective is not to establish causal relationships among organizational variables, but to anticipate workload sufficiently well to support staffing and contingency decisions [14].

The available history covers a short temporal span. Although 122 daily observations do not preclude the estimation of several standard forecasting methods, six months of data provide no repeated annual cycle and limited evidence for reliably identifying persistent seasonality, long-term trends, or structural changes [12,13,15]. The restriction to parsimonious benchmarks is therefore a deliberate design choice rather than a technical necessity imposed by the number of observations. Mean, Naive, and Drift represent three transparent reference structures: a stable historical level, persistence of the most recent observation, and continuation of the average historical change, respectively [13] Section 5.2. Their use establishes a readily interpretable performance threshold against which additional model structure can subsequently be evaluated. This benchmark-oriented design also prioritizes transparency, interpretability, and reproducibility in an operational setting where forecasts must ultimately be updated and translated into workforce-capacity decisions.

Such benchmark-based evaluation is consistent with evidence showing that greater forecasting complexity does not systematically translate into superior out-of-sample performance [16,17]. More recent work likewise emphasizes the value of balancing accuracy, interpretability, computational requirements, and model complexity rather than assuming that the most elaborate specification is necessarily preferable [18]. Simple methods can also have well-defined statistical foundations and remain useful when forecasts must be interpreted, monitored, and updated in operational settings [13,19].

Forecast evaluation must nevertheless distinguish genuine predictive performance from in-sample fit. For time-ordered observations, random partitioning can allow future information to influence model estimation, whereas rolling-origin evaluation preserves temporal order by generating each forecast only from information available at that forecast origin [20,21]. Multiple error measures are also needed because different criteria emphasize different characteristics of forecast performance [22,23]. MAE and RMSE express deviations in the original demand unit, while MASE and sMAPE provide complementary scaled and proportional perspectives.

This distinction is especially relevant for workforce planning because forecast errors expressed in words can be translated directly into productive-capacity equivalents. Scale-dependent errors therefore have an immediate operational interpretation in terms of workload deviation, while their consequences may appear as capacity shortages, unused capacity, backlog, or temporary staffing requirements. Statistical accuracy and operational decision quality are related but not identical, particularly when staffing is discrete and the consequences of overcapacity and undercapacity differ [24].

Although forecasting research has developed extensive methods for predictive evaluation [12,20,24], service-operations and workforce-planning research has focused primarily on demand–capacity alignment and resource decisions [3–8,25,26]. These perspectives are not always connected explicitly in small knowledge-intensive service settings. In particular, less attention has been given to transparent procedures that start from short operational records, evaluate parsimonious benchmarks through temporally valid out-of-sample forecasts, and convert forecast demand and forecast error into workforce-capacity indicators. Translation services provide a suitable setting for examining this connection because workload can be expressed in a measurable operational unit and compared directly with an explicit productivity assumption.

Against this background, the following research question is proposed:

How can parsimonious benchmark models be used to forecast demand and estimate workforce capacity requirements in translation services when only a short historical window is available?

Accordingly, this study analyzes and forecasts the demand of a translation-service team and translates the resulting estimates into workforce-capacity requirements. The empirical analysis uses 122 daily demand observations collected from January to June 2025. Demand is measured as the number of words requested per working day, while workforce capacity is evaluated under an explicit nominal productivity assumption. Mean, Naive, and Drift benchmarks are compared through expanding-window rolling-origin evaluation using complementary scale-dependent, scaled, proportional, and bias-related error measures.

The study makes three contributions. First, it applies transparent benchmark forecasting to a knowledge-intensive service in which forecast demand can be linked directly to a measurable capacity unit. Second, it provides a decision-oriented interpretation of forecasting performance by relating forecast errors and predicted workloads to staffing, utilization, shortage, and contingency-capacity requirements. Third, it develops a reproducible data-to-

decision procedure connecting operational data capture, validation, benchmark evaluation, capacity assessment, managerial action, and subsequent model updating.

The remainder of the paper is organized as follows. Section 2 develops the theoretical background concerning operational data, demand–capacity alignment, parsimonious forecasting benchmarks, and decision-oriented forecast evaluation. Section 3 describes the study design, data construction, capacity assumptions, forecasting methods, rolling-origin evaluation, and demand–capacity analysis. Section 4 presents the empirical results. Section 5 discusses the findings, implications, limitations, and future research directions. Section 6 concludes the study.

2. Theoretical Background

2.1. Operational Data and Predictive Decision Support

Operational decision-making increasingly depends on an organization's ability to transform routinely generated data into consistent and interpretable evidence. However, the availability of data does not by itself guarantee useful analytical results. Forecasting requires observations collected under stable definitions, expressed in comparable units, associated with an appropriate time index, and processed using transparent procedures for identifying missing values, duplicates, and inconsistencies [9,20]. Data preparation is therefore not merely a technical preliminary step but a condition for producing forecasts that can be meaningfully translated into managerial actions.

A distinction must also be made between explanatory and predictive modeling. Explanatory models are primarily intended to describe relationships or evaluate theoretical mechanisms, whereas predictive models aim to generate accurate estimates of unobserved or future outcomes [10,14,27]. A model may consequently offer a plausible explanation of past behavior without providing sufficiently accurate forecasts for operational planning. Because the present study is concerned with estimating future workload and converting it into staffing requirements, its analytical purpose is predictive and decision-oriented rather than causal.

In this context, data management should be understood as the set of procedures that creates a reliable forecasting input. These procedures include defining the unit of analysis, standardizing recording practices, organizing observations chronologically, verifying their completeness, and maintaining consistency between the variables used to represent demand and capacity. Data-analytic processes generate organizational value only when the original records are transformed into variables that are sufficiently reliable, interpretable, and relevant to the decision under consideration [19,22,23]. Related evidence from Ecuadorian SMEs has likewise identified the lack of structured databases and qualified personnel as important constraints on the effective adoption of data-intensive technologies [28].

For translation services, the number of words requested per working day provides an observable workload unit that can subsequently be compared with the effective daily output of the available workforce. The use of a standardized output measure is consistent with service-operations approaches that recommend defining productivity indicators according to the specific nature and delivery process of the service [11]. Nevertheless, the usefulness of word count depends on its consistent recording across the complete observation period and does not eliminate differences in document complexity or processing effort.

The value of predictive analysis ultimately depends on whether its outputs improve operational decisions. Forecasts should therefore not be treated as isolated numerical results. They form part of a broader decision process in which historical information is converted into expected demand, expected demand is translated into resource requirements, and subsequent observations are used to assess whether the forecasting and capacity assumptions remain appropriate. This iterative relationship between data preparation,

model estimation, operational action, and monitoring provides the analytical foundation for the procedure applied in this study.

2.2. Demand–Capacity Alignment in Knowledge-Intensive Services

Capacity management concerns the alignment of available operational resources with the volume of work to be performed over a specified period. In service settings, capacity can be understood as the maximum or effective volume of service that an operational system can deliver within that period, but effective capacity depends not only on the number of available workers but also on productive time, skills, task allocation, workload complexity, and supporting resources [25], Ch. 11, pp. 301–333; [8], Ch. 11, pp. 297–332; [29], pp. 274–309; [30]. Demand represents the volume of service requested during the same period; meaningful demand–capacity assessment therefore requires both variables to be expressed in compatible units and time intervals. In professional services, nominal and effective capacity may differ because meetings, coordination, quality review, interruptions, leave, and heterogeneous task difficulty reduce the time available for direct production.

Balancing demand and capacity is particularly difficult in services because productive time is perishable and generally cannot be stored for later use [1–3,31]. The simultaneity and customer participation associated with many services further limit the ability to separate production from delivery [29,32]. When demand exceeds available capacity, the immediate consequences may include delayed completion, accumulated work, employee overload, or the need to acquire temporary external capacity. Conversely, when capacity remains substantially above demand, organizations incur idle time and underutilize specialized personnel.

Demand variability intensifies this problem. A workforce level based exclusively on average demand may be adequate during typical periods but insufficient when short-term peaks occur. Alternatively, permanent staffing based on the highest historical workload may protect service levels but produce persistent excess capacity during normal periods. Accordingly, capacity planning should distinguish between the core workforce required to serve regular demand and flexible or contingency capacity used to absorb unusual workload increases [4–6].

Operations-management theory also shows that variability has nonlinear consequences as resource utilization increases. When a system operates close to its effective capacity, relatively small workload fluctuations can produce large increases in queues, delays, and accumulated work [7], Chs. 8–9. Thus, a high average utilization rate should not automatically be interpreted as operational efficiency, particularly when the arrival and processing of requests are irregular.

This distinction is especially relevant in knowledge-intensive services. Their capacity is embedded primarily in specialized professionals and cannot always be expanded immediately. Recruitment and training may require substantial time, while outsourcing or temporary contracting may introduce differences in cost, availability, coordination, confidentiality, and service quality. Competency-based workforce management can further support this alignment by connecting workforce capabilities with organizational processes and performance requirements [33]. Workforce capacity should therefore be interpreted as an effective operational resource rather than merely as the total number of employees assigned to a unit.

Translation services illustrate this relationship clearly. Incoming documents create a measurable workload, while translators provide a finite amount of productive capacity during each working period. If demand and productivity are expressed using the same unit, forecast demand can be converted into an estimated staffing requirement. However, this conversion necessarily depends on an assumed productivity rate. The resulting staffing

estimate should therefore be interpreted as conditional on that assumption rather than as an exact or universally applicable number.

The direction of translation, document complexity, terminology, formatting requirements, revision activities, deadlines, interruptions, and nontranslation responsibilities may all affect effective output. Consequently, a nominal productivity standard is useful for planning but does not capture the full variability of individual or team capacity. A robust demand–capacity analysis should therefore distinguish between expected staffing needs and the additional flexibility required to protect operations against productivity variation and temporary demand peaks.

2.3. Parsimonious Benchmarks Under Limited Historical Data

Time-series forecasting uses observations recorded sequentially to characterize temporal behavior and estimate future values. Classical time-series analysis emphasizes that the appropriate model depends on the stochastic structure of the data, including its level, dependence, trend, seasonality, changing variance, and irregular variation ([13], Chs. 2–3); [15,34]. The selected model should therefore reflect both the characteristics of the series and the quantity of historical information available [20,34].

Long series may allow the estimation of trend, seasonal, cyclical, intervention, and other systematic components. Short series, by contrast, limit the number of patterns that can be identified reliably and increase the risk that a complex model will reproduce random variation rather than a stable demand process. Standard time-series methodology consequently requires that the complexity of the specification be supported by the length and informational content of the observed series ([13], Chs. 2–3); [15,34].

The present context is characterized by a short temporal observation window rather than by an insufficient number of observations for model estimation. The 122 daily observations could support the estimation of other standard forecasting approaches, including exponential-smoothing or ARIMA-type models. However, the six-month span does not contain repeated annual cycles and provides limited evidence for reliably identifying persistent seasonal, long-term, or structural patterns. The restriction to parsimonious benchmarks was therefore a deliberate analytical choice intended to establish transparent reference performance while avoiding additional structural assumptions that were not central to the purpose of the study.

Three common benchmarks are particularly relevant [34]. The Mean method assigns the historical average of the series to every future period and is appropriate as a baseline when no persistent trend or seasonality can be established. The Naive method uses the most recent observation as the forecast for subsequent periods and represents a random-walk benchmark. The Drift method extends the Naive approach by projecting the average change observed between the first and last observations across the forecast horizon.

These models embody different assumptions. The Mean method assumes that observations fluctuate around a relatively stable level. The Naive method gives maximum importance to the latest value and is therefore more responsive to recent changes, but it can also transmit an unusually high or low final observation throughout the forecast horizon. The Drift method assumes that the average historical rate of change will continue, making its long-horizon forecasts particularly sensitive to the first and last observations. Comparing the three methods helps determine whether the available evidence supports a stable-level, persistence, or linear-change representation of demand.

The use of benchmark models does not imply that simple methods are intrinsically superior. Their principal purpose is to establish a credible performance threshold against which additional complexity can be evaluated. Large forecasting studies have demonstrated that sophisticated models do not consistently outperform parsimonious alternatives

and that relatively simple statistical methods remain competitive across many types of series [16–18]. Simplicity is especially valuable when data are limited, model outputs must be communicated to operational managers, and forecasts must be updated without highly specialized computational infrastructure.

Moving averages and exponential smoothing may complement this analysis by describing local behavior or reducing short-term noise. However, descriptive smoothing should not be confused with formal forecast evaluation. A smoothed representation can facilitate visual interpretation of the series, but a model should be selected on the basis of forecasts generated without access to the observations being predicted. Accordingly, the empirical comparison should clearly distinguish exploratory visualization, fitted values, and genuine out-of-sample forecasts [9,34].

A recent historical and methodological review likewise identifies benchmark forecasting methods as essential reference tools and emphasizes simplicity, parsimony, computational efficiency, transparency, and out-of-sample reliability as core principles underlying their continued use in forecasting practice [35].

2.4. Forecast Evaluation and Decision-Oriented Model Selection

Forecast accuracy must be evaluated using observations that were not used to generate the corresponding predictions. Errors calculated from fitted values may underestimate actual forecasting uncertainty because model parameters can incorporate information from the complete series. For time-ordered data, evaluation procedures must also preserve chronological order. A holdout period or rolling-origin evaluation is therefore preferable to random partitioning, which can introduce information from future observations into model estimation [9,34].

No single error measure fully characterizes forecasting performance. Mean error (ME) indicates systematic directional bias, with positive and negative errors potentially cancelling one another. Mean absolute error (MAE) expresses the average magnitude of deviations in the original unit of the series and is readily interpretable in operational terms. Mean squared error (MSE) and root mean squared error (RMSE) give greater weight to large deviations, making them useful when substantial forecasting errors are especially costly.

Scaled and percentage measures provide complementary perspectives. Mean absolute scaled error (MASE) compares absolute forecast error with the error of a reference benchmark and remains interpretable across series with different scales. Symmetric mean absolute percentage error (sMAPE) expresses proportional deviations but, like other percentage measures, requires cautious interpretation when observed values are zero or close to zero ([13] Section 5.8) [22,23]. The revised empirical analysis therefore prioritizes MAE, RMSE, MASE, and sMAPE rather than relying on conventional MAPE.

Different measures may generate different model rankings because they impose different penalties on the distribution of forecast errors [22,23]. Conflicting rankings should not be treated as an analytical failure. Rather, it demonstrates that model selection requires an explicit criterion connected to the intended use of the forecast.

For workforce capacity planning, errors expressed in the original demand unit have a direct managerial interpretation. In translation services, an absolute error measured in words can be converted into productive time and, subsequently, into excess or insufficient staffing capacity. RMSE may be particularly informative when large underestimates create severe workload pressure, whereas MAE describes the typical magnitude of the staffing deviation. MASE indicates whether a model improves upon a simple reference, while sMAPE provides supplementary proportional information.

This relationship supports a decision-oriented approach to model selection. Statistical accuracy and decision quality are related but not identical. A model with the lowest value for a particular error measure may not necessarily generate the most appropriate operational decision, especially when staffing requirements are discrete and when the costs of overcapacity and undercapacity differ [24]. Model evaluation should therefore consider how each forecast affects the estimated workforce requirement, the frequency and magnitude of capacity shortages, and the need for contingency resources.

The theoretical framework developed in this section consequently comprises three connected stages. First, operational data must be standardized into a consistent demand series. Second, parsimonious benchmarks must be evaluated through a temporally valid procedure and multiple complementary error measures. Third, forecast demand must be converted into capacity requirements using an explicit productivity assumption. This structure links forecasting performance to its operational consequences and provides the conceptual basis for the methodological procedure developed in the following section.

3. Materials and Methods

3.1. Study Design and Operational Setting

This research was designed as an applied, quantitative, retrospective longitudinal single-case study [36], Ch. 2. Its purpose was predictive and decision-oriented: to examine the historical demand of a translation service, compare parsimonious benchmark forecasting models, and convert the resulting forecasts into workforce capacity requirements. The study was not intended to establish causal relationships or test the effects of organizational interventions.

The empirical setting was an internal translation team composed of seven professionals: one coordinator and six translators. Three translators were assigned primarily to English-language services and three to Spanish-language services. The team processed educational materials, institutional communications, contracts, presentations, and other organizational documents. Requests were received mainly through Trello and email and were distributed among translators by the team coordinator.

The working day was adopted as the primary unit of analysis because both incoming demand and productive capacity were managed on a daily basis. All forecasting analyses were therefore performed using chronologically ordered working-day observations rather than monthly averages. Monthly summaries were calculated only for descriptive reporting and managerial interpretation.

3.2. Data Sources and Construction of the Demand Series

The study covered the period from 5 January to 30 June 2025. The analytical dataset contained 122 working-day observations. The complete daily demand dataset is provided in Supplementary Table S1. Demand records were obtained from the operational channels through which translation requests were received, principally Trello and email, and were consolidated with the assistance of the coordinator responsible for workload allocation.

A standardized spreadsheet was created to organize the historical records. For each working day, the requests received through the different operational channels were consolidated and expressed as the total number of words requested. The resulting series was defined as

$$y_t = \text{number of words requested on working day } t, \qquad t = 1, \ldots, 122$$

Word count was selected as the workload unit because the documents differed substantially in format, page length, and presentation. Page counts were not considered

comparable across Word files, spreadsheets, presentations, contracts, and other materials. Expressing demand in words provided a common quantitative unit that could be compared with translator productivity.

The chronological sequence was checked before analysis. Dates corresponding to weekends and organizational nonworking days were not incorporated because neither regular demand processing nor standard productive capacity was expected on those dates. Recorded zero-demand working days were retained as valid observations rather than treated as missing values. No interpolation was applied to the demand series.

These validation procedures were intended to preserve the accuracy, completeness, consistency, and temporal interpretability of the analytical dataset. These dimensions are central to data quality because deficiencies in source records or transformation procedures may propagate into the resulting information products and affect subsequent decisions [37].

The available analytical file contained an aggregate daily word count and did not provide independent demand series for English and Spanish. Consequently, all forecasting and workforce-capacity estimates in the present study refer to the combined translation workload and should be interpreted at the team level. The available data do not support allocation of the resulting headcount requirement between language directions or subteams.

3.3. Measurement of Workforce Capacity

The nominal productive capacity of one translator was established as an organization-specific planning assumption based on a one-week workplace observation combined with a structured discussion involving the researchers, translators, and team coordination. During the observation period, the volume of translated words completed over eight-hour working days was examined and interpreted together with the team's operational experience. On this basis, a nominal value of 2000 translated words per translator per working day was adopted. This value was not taken from an external industry standard and should not be interpreted as a universal translation-productivity rate.

The parameter represents nominal productive output for an available translator rather than the effective capacity of the workforce as a whole. Effective team capacity may be lower because revision and quality-assurance activities, meetings, coordination, administrative duties, interruptions, annual or sick leave, and other non-translation responsibilities reduce either productive time or the number of translators available on a particular day. Organizational nonworking days were excluded from the analytical series, but temporary variation in staff availability was not modeled separately. Consequently, all staffing estimates are conditional on the stated productivity and availability assumptions.

Let (c) represent the assumed productive capacity of one translator:

$$c = 2000 \text{ words per translator per working day}$$

For an observed or forecast demand value (d_t), the corresponding workforce requirement was calculated as

$$N_t = \text{ceil}\left(\frac{d_t}{c}\right)$$

where $\text{ceil}\left(\frac{d_t}{c}\right)$ denotes rounding upward to the next whole translator. Upward rounding was applied because partial-staffing units cannot independently cover the full daily workload.

Because the capacity estimate was based on a short observational period, a sensitivity analysis was specified using three productivity scenarios:

- Conservative capacity: 1600 words per translator per day;
- Nominal capacity: 2000 words per translator per day;

- Favorable capacity: 2400 words per translator per day.

These scenarios represent a variation of ±20% around the nominal capacity and allow the stability of the staffing conclusions to be assessed under plausible productivity differences.

3.4. Exploratory Time-Series Analysis

The daily demand series was first examined descriptively to characterize its level, variability, extreme observations, and temporal evolution. The analysis included the mean, median, standard deviation, interquartile range, minimum, maximum, and coefficient of variation. The number and proportion of zero-demand days were also calculated.

A five-working-day moving average was used to visualize weekly changes in the local demand level. Simple exponential smoothing with a smoothing parameter ($\alpha = 0.10$) was also used as a descriptive representation of the series' longer-term level. These smoothed series were employed exclusively for exploratory visualization and were not used to select the forecasting model.

No annual seasonality or long-term cyclical structure was estimated. Although the dataset contained 122 daily observations, it covered only six months and therefore did not include repeated annual cycles. Apparent peaks or recurring movements within the observation period were interpreted as short-term variation rather than evidence of established seasonality.

3.5. Parsimonious Benchmark Models

Three benchmark forecasting methods were evaluated: Mean, Naive, and Drift [13], Section 5.2. Their selection was deliberate and reflected the benchmark-oriented purpose of the study rather than an inability to estimate more structured models with the available number of observations. The three methods provide distinct and readily interpretable reference representations of demand: Mean represents a stable historical level, Naive represents persistence of the most recent observation, and Drift represents continuation of the average historical rate of change. They also require minimal parameterization and provide transparent forecasts that can be readily interpreted in terms of workforce-capacity consequences. The analysis therefore evaluates whether these basic representations provide an adequate operational reference before considering additional model complexity.

For a training series ($\hat{y}_{T+h|T}$), the Mean method assigns the historical average to every future period:

$$\hat{y}_{T+h|T} = \frac{1}{T}\sum_{t=1}^{T} y_t$$

The Naive method assigns the most recently observed value to all subsequent periods:

$$\hat{y}_{T+h|T} = y_T$$

This method represents a random-walk benchmark and gives maximum weight to the latest available observation.

The Drift method extends the Naive forecast by continuing the average change observed between the first and last values of the training series:

$$\hat{y}_{T+h|T} = y_T + h\left(\frac{y_T - y_1}{T-1}\right)$$

Because the Drift forecast can become negative when the estimated direction is downward, forecasts below zero were truncated at zero, as negative translation demand has no operational interpretation.

3.6. Rolling-Origin Forecast Evaluation

Model performance was assessed through expanding-window rolling-origin evaluation rather than through in-sample fitted errors ([13,20] Sections 5.8 and 5.10) [21]. This approach preserved the chronological structure of the data and ensured that each forecast was generated using only observations available before the period being predicted.

The first 60 working-day observations were used as the initial estimation window. At the first forecast origin, each model was estimated using days 1–60 and used to predict day 61. The estimation window was then expanded by one observation, and the procedure was repeated until the end of the series. This produced 62 sequential one-step-ahead forecast errors [21].

One-working-day-ahead forecasts constituted the primary evaluation horizon because short-term workload allocation and contingency decisions were made at the daily operational level. Five-working-day and twenty-working-day horizons were also examined as sensitivity analyses, approximating weekly and monthly planning decisions, respectively. At every origin and horizon, the forecasting model was re-estimated using only the historical information available at that point.

Forecast errors were defined as

$$e_t = y_t - \hat{y}_t$$

where (y_t) is observed demand and $(\hat{y}_t)$ is forecast demand.

Performance was evaluated using complementary measures [13] Section 5.8, [20,22,23]:

- Mean error (ME), to assess systematic overforecasting or underforecasting;
- Mean absolute error (MAE), to quantify the typical error in words;
- Root mean squared error (RMSE), to give greater weight to large errors;
- Mean absolute scaled error (MASE), to provide a scale-free comparison relative to a Naive benchmark;
- Symmetric mean absolute percentage error (sMAPE), to describe proportional error while reducing some of the asymmetry of conventional percentage measures.

Conventional MAPE was not used as the primary selection criterion because the series contained zero and near-zero demand observations, for which percentage errors are undefined or disproportionately large. When reported as supplementary information, MAPE was calculated only for observations with strictly positive actual demand, and the number of excluded zero observations was stated explicitly.

The model with the lowest one-step-ahead MAE was selected as the primary operational benchmark because MAE measures forecast deviations in words and can be translated directly into productive time and staffing requirements. RMSE was used as a secondary criterion to assess sensitivity to large errors, while ME was examined to identify systematic bias. The weekly and monthly horizon results were used to determine whether the model ranking remained stable as the forecast horizon increased.

To facilitate operational interpretation, the scale-dependent forecast errors were also expressed as translator-day equivalents by dividing the corresponding error in words by the nominal productive capacity of 2000 words per translator per working day. Thus, an absolute forecast deviation of 2000 words represents one translator-day of nominal productive capacity. These equivalents indicate the magnitude of the workload deviation in capacity units rather than a discrete staffing change, because actual headcount decisions require upward rounding and also depend on delivery deadlines, acceptable backlog, and the direction of the forecast error. Positive ME indicates average underforecasting and therefore potential capacity pressure, whereas negative ME indicates average overforecasting and potential unused capacity.

3.7. Final Forecasts and Forecast Uncertainty

After model comparison, the selected benchmark was re-estimated using all 122 daily observations. Forecasts were then generated for the subsequent planning period. Point forecasts were accompanied by 80% and 95% prediction intervals to communicate the uncertainty associated with future demand and to distinguish expected demand from plausible higher- and lower-demand outcomes [38].

The long-horizon forecasts for July–December 2025 were treated as planning scenarios rather than as externally validated predictions. Their interpretation was necessarily cautious because the forecast horizon was approximately as long as the available historical period. Monthly values were obtained by summarizing the corresponding working-day forecasts; the forecasting models themselves were not estimated from only six monthly observations.

For capacity planning, the point forecast represented the expected or baseline staffing requirement. The upper prediction limit was used to assess contingency capacity, reflecting the number of translators potentially required under a high-demand but statistically plausible scenario.

3.8. Demand–Capacity Alignment

Observed and forecast demand were converted into operational capacity indicators. For each working day, the analysis is calculated:

1. Required number of translators;
2. Available productive capacity;
3. Utilization ratio;
4. Capacity shortage in words;
5. Unused capacity in words.

For a workforce of (A) available translators, total available daily capacity was defined as

$$C_t = A \times c$$

The utilization ratio was calculated as

$$U_t = \frac{d_t}{C_t}$$

Values above 1 indicated that demand exceeded nominal capacity, whereas values below 1 indicated available capacity greater than recorded demand.

Daily capacity shortage and unused capacity were calculated as

$$S_t = \max(d_t - C_t, 0)$$

and

$$E_t = \max(C_t - d_t, 0)$$

respectively.

The analysis summarized the number and percentage of days with insufficient capacity, total accumulated shortage, maximum daily shortage, average utilization, and the staffing requirement associated with typical and peak demand. The same calculations were repeated under the conservative, nominal, and favorable productivity scenarios.

3.9. Data-to-Decision Procedure

The analytical process was organized into five connected stages, as summarized in Table 1. Unlike the original administrative description, each stage was linked to a repro-

ducible set of activities and a clearly defined analytical or managerial output. The procedure progresses from standardized data capture and validation to benchmark forecasting, capacity assessment, decision support, and continuous model updating. The monitoring stage also feeds newly observed demand and forecast errors back into the data-capture process, allowing the forecasting models and productivity assumptions to be periodically revised.

Table 1. Data-to-decision procedure for translation-demand forecasting and capacity planning.

Stage	Activities	Main Output
1. Data capture	Collect requests from Trello and email; record the date and number of words using a standardized format.	Chronologically ordered operational records.
2. Data validation and preparation	Verify dates and units; distinguish zero demand from missing records; aggregate requests by working day; document data limitations.	Validated daily demand series.
3. Forecasting and evaluation	Explore the series; estimate Mean, Naive, and Drift benchmarks; apply rolling-origin evaluation; compare forecast errors and uncertainty.	Selected benchmark and prediction intervals.
4. Capacity translation and decision support	Convert observed and forecast demand into staffing, utilization, shortage, and excess-capacity indicators; assess productivity scenarios.	Baseline and contingency workforce requirements.
5. Monitoring and model updating	Update the database as new requests arrive; compare forecasts with realized demand; review the productivity assumption; periodically re-estimate the models.	Continuous feedback and updated capacity decisions.

Figure 1 summarizes the iterative relationship among these five stages. The monitoring stage feeds new observations and observed forecast errors back into the data-capture stage, enabling periodic revision of both the forecasting model and the capacity assumptions.

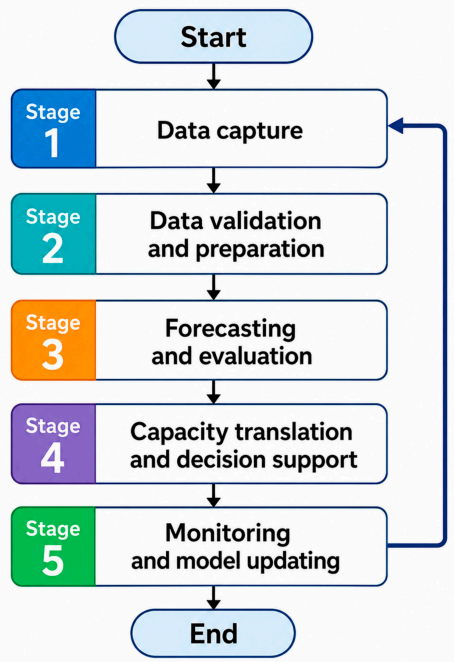


Figure 1. Data-to-decision workflow for forecasting translation demand and planning workforce capacity.

3.10. Software, Reproducibility, and Data Governance

Microsoft Excel was used only for the initial consolidation and inspection of the operational records. Data preparation, descriptive analysis, rolling-origin evaluation, forecast generation, accuracy measurement, sensitivity analysis, and graphical presentation were conducted in R version 4.3.2 using the packages forecast (v8.21.1), fable (v0.3.3), tsibble (v1.1.4), dplyr (v1.1.4), and ggplot2 (v3.4.4).

The analysis used routine organizational workload records and did not retain translators' names, individual productivity scores, customer identities, document contents, or other direct personal or commercially sensitive identifiers. Results were reported at the aggregate team level. The analytical code and a de-identified version of the daily demand series may be made available subject to organizational authorization and applicable confidentiality requirements.

4. Results

4.1. Descriptive Characteristics of Daily Demand

The final time series contained 122 working-day observations from 5 January to 30 June 2025, and the complete observation-level dataset is reported in Supplementary Table S1. Mean daily demand was 5472.12 words, whereas the median was 4373.09 words. The difference between these values reflects the influence of several high-demand observations. Daily demand had a standard deviation of 4091.44 words and an interquartile range of 4632.40 words, ranging from 0 to 19,686 words. The coefficient of variation was 74.77%, indicating substantial variability relative to the average workload. Two working days recorded zero demand.

Figure 2 presents the complete daily series together with the five-working-day moving average and the simple exponential smoothing series with $\alpha = 0.10$.

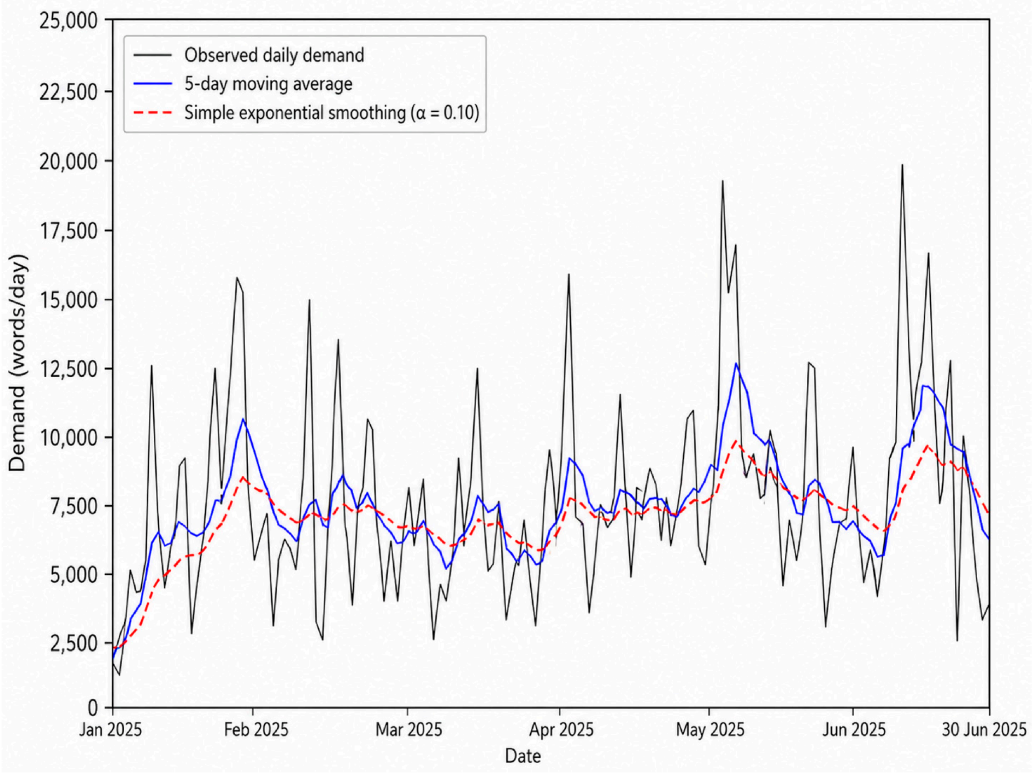


Figure 2. Daily translation demand with five-day moving average and simple exponential smoothing (January–June 2025). Smoothed series are shown for exploratory visualization only.

Daily demand fluctuated markedly throughout the observation period. The five-day moving average ranged from 2749.40 to 12,272.40 words, demonstrating that substantial workload changes persisted even after short-term daily variation was smoothed. The highest daily demand was recorded on 16 June, with 19,686 words. Other substantial peaks occurred on 10 May (18,335 words), 12 May (16,209 words), 21 June (15,786.50 words), 30 January (15,006 words), and 31 January (14,795.50 words). These peaks were not distributed at regular intervals and therefore did not provide sufficient evidence of a recurring seasonal pattern.

Monthly descriptive results are reported in Table 2. May had the highest average demand, at 6947.23 words per working day, followed by January, June, April, February, and March. June had the highest observed daily value and the greatest within-month standard deviation. Because the six-month observation window did not contain repeated annual cycles, the monthly differences were treated as descriptive variations rather than as evidence of seasonality.

Table 2. Descriptive summary of daily translation demand by month.

Month	Working Days	Mean	Median	SD	Minimum	Maximum	Translators Required at the Monthly Mean
January	19	5872.40	4400.17	4836.85	0.00	15,006.00	3
February	19	4987.39	4270.00	3857.64	391.00	14,571.00	3
March	23	4030.62	3502.75	2323.23	597.00	10,312.00	3
April	18	5458.61	4753.25	2511.88	2216.00	11,963.00	3
May	22	6947.23	7118.00	4594.47	864.00	18,335.00	4
June	21	5593.52	3765.50	5309.25	159.00	19,686.00	3
Overall	122	5472.12	4373.09	4091.44	0.00	19,686.00	3

Note: Staffing requirements are based on the nominal capacity of 2000 words per translator per working day and are rounded upward to the next whole translator. All staffing estimates refer to the aggregate translation workload.

At the nominal productivity level, the overall average demand corresponded to 2.74 translator-equivalents and therefore to three translators after upward rounding. As a descriptive observation, the May monthly mean corresponded to four translator-equivalents after upward rounding, whereas the other monthly means corresponded to three. This monthly contrast is reported for managerial interpretation only and should not be interpreted as evidence of a persistent change in the underlying demand level.

4.2. One-Step-Ahead Benchmark Performance

The Mean, Naive, and Drift benchmarks were first compared through expanding-window rolling-origin evaluation. The first 60 observations constituted the initial estimation window, and each model generated a one-working-day-ahead forecast before the estimation sample was expanded by one observation. This procedure produced 62 genuine out-of-sample forecast errors for each model. Table 3 shows that the Mean method provided the best overall predictive performance.

Its MAE was 3221.66 words, compared with 4589.02 for Naive and 4618.67 for Drift. Its RMSE was also substantially lower, at 4424.81 words, whereas Naive and Drift produced RMSE values of 5831.89 and 5863.91 words, respectively.

The Mean benchmark was the only model with a MASE below 1, indicating that it improved upon the in-sample Naive scaling benchmark. Its MASE of 0.845 represented an approximately 15.5% reduction in mean absolute scaled error relative to that reference. It also produced the lowest sMAPE, at 58.14%.

Table 3. One-working-day-ahead rolling-origin forecast performance.

Model	Forecasts	ME	MAE	RMSE	MASE	sMAPE (%)
Naive	62	−74.98	4589.02	5831.89	1.203	80.42
Mean	62	683.16	3221.66	4424.81	0.845	58.14
Drift	62	−144.72	4618.67	5863.91	1.211	80.53

Note: Forecast error was defined as observed minus forecast demand. Positive ME therefore indicates average underforecasting, whereas negative ME indicates average overforecasting. MASE was scaled using the mean absolute one-step Naive error in the initial 60-day estimation window.

The ME results revealed differences in forecast direction. The Mean model had a positive ME of 683.16 words, indicating that it underestimated demand on average during the evaluation period. Naive and Drift had small negative mean errors, indicating slight average overforecasting. Nevertheless, their lower directional bias did not compensate for their substantially larger absolute errors.

Expressed in operational capacity units, the Mean model's MAE of 3221.66 words was equivalent to approximately 1.61 translator-days of nominal productive capacity, while its RMSE of 4424.81 words corresponded to approximately 2.21 translator-days. Its positive ME of 683.16 words represented an average underforecasting bias equivalent to approximately 0.34 translator-days. By comparison, the MAE values of the Naive and Drift models corresponded to approximately 2.29 and 2.31 translator-days, respectively, while their RMSE values corresponded to approximately 2.92 and 2.93 translator-days. These conversions reinforce the operational advantage of the Mean benchmark among the models evaluated: its typical and large forecast deviations represented less productive capacity than those of the alternative benchmarks.

Figure 3 presents the observed evaluation-period demand and the sequential one-step-ahead forecasts.

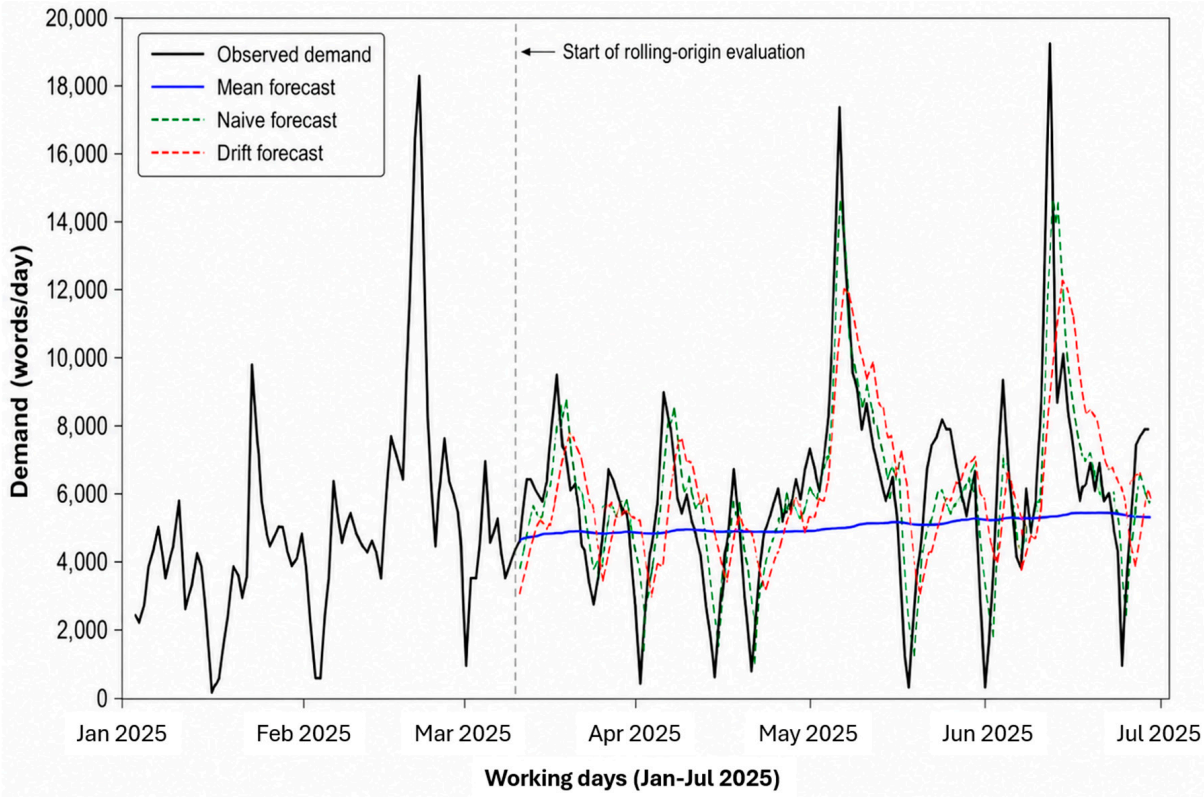


Figure 3. Observed demand and one-step-ahead rolling-origin forecasts generated by the Mean, Naive, and Drift benchmarks.

The Mean forecast changed gradually as the estimation window expanded, whereas Naive and Drift reacted more strongly to the latest observation. This greater responsiveness caused both models to transmit extreme daily values into subsequent forecasts, increasing their absolute errors in the highly variable series.

These error equivalents characterize the magnitude of forecast uncertainty in productive-capacity units; their actual staffing consequences depend on the available workforce, realized demand, and operational constraints. Section 4.5 therefore complements the forecast-accuracy comparison by examining the frequency and magnitude with which observed demand exceeded the forecast-derived three-translator baseline capacity, together with the temporary aggregate capacity implied by the forecast distribution.

4.3. Forecast-Horizon Sensitivity

The performance comparison was repeated at horizons of five and twenty working days to examine whether model rankings changed under approximate weekly and monthly planning conditions. The Mean method maintained the lowest MAE, RMSE, MASE, and sMAPE at both horizons. As shown in Table 4, the Mean method maintained the lowest MAE, RMSE, MASE, and sMAPE at both horizons.

Table 4. Forecast performance across alternative planning horizons.

Horizon	Model	Forecasts	ME	MAE	RMSE	MASE	sMAPE (%)
1 working day	Naive	62	−74.98	4589.02	5831.89	1.203	80.42
	Mean	62	683.16	3221.66	4424.81	0.845	58.14
	Drift	62	−144.72	4618.67	5863.91	1.211	80.53
5 working days	Naive	58	−94.18	4699.80	6442.95	1.233	75.26
	Mean	58	770.92	3289.65	4529.51	0.863	58.19
	Drift	58	−452.96	4892.51	6623.77	1.283	76.31
20 working days	Naive	43	191.09	4608.33	6424.38	1.209	74.45
	Mean	43	1089.99	3758.68	5028.54	0.986	66.40
	Drift	43	−1354.06	5055.15	7117.23	1.326	75.84

The Mean model's MAE increased from 3221.66 words at the one-day horizon to 3758.68 words at the twenty-day horizon. Its MASE also approached 1 as the horizon increased, rising from 0.845 to 0.986. Thus, although the Mean benchmark remained the most accurate of the three models, its relative advantage decreased for longer-term forecasts.

Drift exhibited the largest deterioration as the horizon increased. Its RMSE rose from 5863.91 words at one day to 7117.23 words at twenty days, and its negative ME increased in magnitude to −1354.06 words. This result indicates that extending the historical average change over longer horizons produced increasingly high forecasts relative to realized demand during the evaluation period.

The consistency of the ranking across the three horizons supported the selection of the Mean method as the primary operational benchmark. However, the magnitude of all error measures also indicated that substantial uncertainty remained, even for the best-performing model.

4.4. Final Demand Forecasts

After model evaluation, the Mean method was re-estimated using all 122 daily observations. The resulting point forecast was 5472.12 words per working day, corresponding to the arithmetic mean of the complete daily demand series. Because the Mean benchmark assumes a stable demand level, this point forecast remained constant across the July–December 2025 planning horizon.

The 80% model-based prediction interval ranged from 178.38 to 10,765.85 words per working day. The untruncated 95% lower limit was negative and was therefore set to zero because demand cannot take negative values; the resulting 95% interval ranged from 0 to 13,605.32 words. Figure 4 presents the final Mean benchmark forecast together with the corresponding 80% and 95% prediction intervals for July–December 2025.

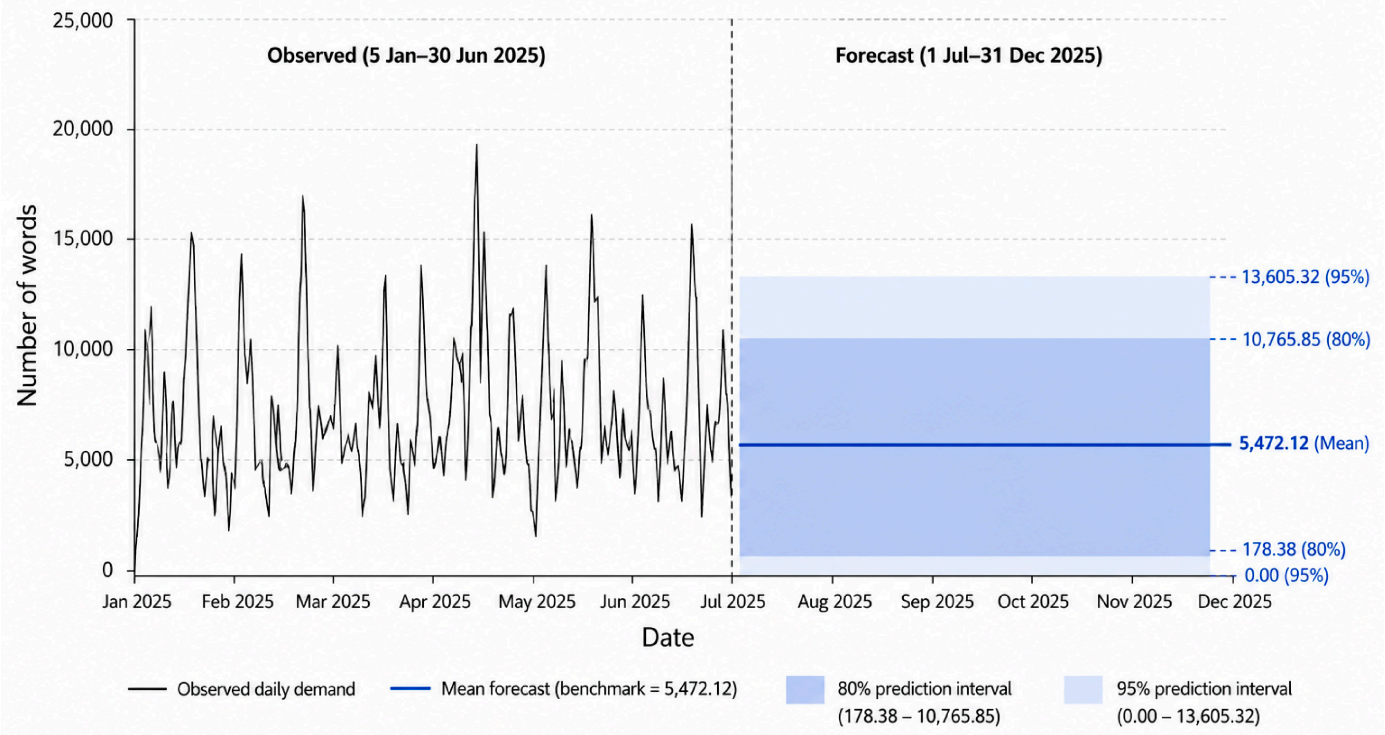


Figure 4. Final Mean benchmark forecast and 80% and 95% prediction intervals for July–December 2025. The Mean benchmark forecast uses the historical average daily demand (5472.12 words) as a constant forecast for July-December 2025. Shaded areas represent 80% and 95% prediction intervals.

The width of the prediction intervals reflected the large historical variability in daily demand. Consequently, the constant point forecast represented an expected workload level rather than a guarantee that daily demand would remain close to 5472 words. The upper limits were used in the subsequent capacity assessment to estimate potential contingency requirements.

4.5. Workforce Capacity Requirements

At the nominal productivity assumption of 2000 words per translator per working day, the final point forecast of 5472.12 words corresponded to 3 translator-equivalents:

$$\frac{5472.12}{2000} = 2.73606 \approx 3 \text{ translators}$$

After upward rounding, this implied a baseline requirement of three translators for the aggregate daily workload. Because the demand series does not distinguish between English and Spanish requests, this estimate cannot be allocated between language-specific subteams.

The 80% upper prediction limit corresponded to 5.38 translator-equivalents and therefore to six translators for the aggregate workload, while the 95% upper prediction limit corresponded to 6.80 translator-equivalents and therefore to seven. These figures should not be interpreted as permanent staffing prescriptions; rather, they quantify the temporary aggregate capacity associated with unusually high but statistically plausible daily demand.

The historical data further showed that the forecast-derived baseline of three translators would not have covered all observed daily workloads. Under the nominal productivity assumption, three translators provide an aggregate daily capacity of 6000 words. Observed demand exceeded this baseline capacity on 42 of the 122 working days, corresponding to 34.43% of the observation period.

Under same-day processing, the maximum observed demand of 19,686 words would have required ten translators at the nominal productivity level. This value should not be interpreted as a permanent staffing requirement because translation assignments may be distributed across several working days when delivery deadlines permit. It does, however, demonstrate the potential magnitude of short-term workload pressure.

Table 5 presents the sensitivity of the staffing conclusions to alternative productivity assumptions.

Table 5. Workforce-capacity sensitivity under alternative individual productivity assumptions.

Individual Capacity (Words/Day)	Capacity of Three Translators	Point-Forecast Requirement	Days Exceeding Three-Translator Capacity	Average Utilization	Maximum Observed Requirement	Requirement at 80% Upper Limit	Requirement at 95% Upper Limit
1600	4800	4	52 (42.62%)	114.00%	13	7	9
2000	6000	3	42 (34.43%)	91.20%	10	6	7
2400	7200	3	31 (25.41%)	76.00%	9	5	6

Under the conservative productivity assumption of 1600 words per day, the point forecast required four translators, and three-translator capacity was exceeded on 42.62% of observed working days. Under the favorable assumption of 2400 words per day, the point forecast still required three translators after upward rounding, but the frequency of historical shortages fell to 25.41%.

These results demonstrate that the baseline headcount is sensitive to the assumed productivity rate: a 20% reduction from the nominal value changes the point-forecast requirement from three to four translators, whereas a 20% increase does not reduce the rounded requirement below three.

Overall, the capacity analysis showed that three translators were sufficient for the expected workload under the nominal productivity assumption but did not provide reliable protection against daily demand peaks. The results therefore distinguished between a baseline requirement, represented by the point forecast, and a contingency requirement, represented by the upper forecast limits and the observed frequency of short-term capacity shortages.

5. Discussion

5.1. Main Findings

This study examined whether parsimonious benchmark models could support demand forecasting and workforce capacity planning in a translation service characterized by limited historical data and substantial daily workload variability. Three main findings emerged.

First, translation demand fluctuated considerably throughout the six-month observation period. Mean daily demand was 5472.12 words, while the standard deviation reached 4091.44 words and the coefficient of variation was 74.77%. The difference between the mean and median, together with several daily peaks above 15,000 words, shows that workload was not concentrated closely around a stable central level. This pattern is consistent with the general characteristics of service operations, in which productive capacity cannot be stored, and short-term demand fluctuations are transmitted directly to workloads, queues, delays, and resource utilization [1–3,7,8,25,26].

The magnitude of this variability is operationally important because the team depended on a small number of specialized professionals. Operations-management theory indicates that variability becomes increasingly disruptive as utilization approaches available capacity, since even moderate fluctuations can produce disproportionate increases in waiting time and accumulated work [7], Chs. 8–9. Accordingly, the average demand level alone was insufficient to characterize the capacity pressure experienced by the translation team.

Second, the Mean benchmark consistently outperformed the Naive and Drift methods under expanding-window rolling-origin evaluation. It produced the lowest MAE, RMSE, MASE, and sMAPE at one-, five-, and twenty-working-day horizons. Its one-day MASE of 0.845 indicates an approximately 15.5% reduction in scaled absolute error relative to the Naive reference. The stability of this ranking across horizons provides a stronger basis for model selection than an in-sample comparison or a single holdout forecast ([13], Sections 5.8 and 5.10) [20,21].

This result does not establish the Mean method as the best forecasting model for this demand series beyond the set of benchmarks evaluated here. Rather, it indicates that, among the three benchmark representations examined, the observed data did not contain a sufficiently persistent recent level or linear direction for the Naive and Drift methods to exploit. The Naive method transmitted unusually high or low observations into the following forecast, while Drift became increasingly sensitive to the average change between the first and most recent observations as the forecast horizon increased. By contrast, the Mean model attenuated the influence of individual extremes and produced a more stable representation of the demand level.

The finding is consistent with evidence showing that greater complexity and responsiveness do not automatically improve forecasting performance [16–18]. Parsimonious models may remain competitive when historical information is limited, the signal-to-noise ratio is low, and persistent seasonal or trend structures cannot be identified. The results also reinforce the methodological importance of evaluating predictions using observations that were unavailable when each forecast was produced [13,20,21].

The operational interpretation of the forecast errors further clarifies the importance of this model comparison. At the nominal productivity level, the Mean benchmark's one-day MAE represented approximately 1.61 translator-days of productive capacity, compared with more than 2.29 translator-days for both Naive and Drift. This does not mean that 1.61 additional translators would routinely be required; rather, it quantifies the typical forecast deviation in the same capacity unit used for subsequent staffing decisions.

Third, translating forecast demand into workforce requirements revealed a clear distinction between average and peak capacity. Under the nominal productivity assumption of 2000 words per translator per working day, the final Mean forecast corresponded to 2.74 translator-equivalents and therefore supported a baseline requirement of three translators. However, demand exceeded the nominal capacity of three translators on 42 of 122 working days, equivalent to 34.43% of the observation period. Average demand and average staffing requirements therefore did not indicate complete demand–capacity alignment.

The monthly summaries provide an illustrative rather than structural perspective on this capacity pressure. Under the nominal productivity assumption, the May mean of 6947.23 words per working day corresponded to four translators after upward rounding, compared with three for the overall daily mean. Because these monthly summaries were not used for model estimation and the six-month observation window does not establish a recurring monthly pattern, the May value should not be interpreted as evidence of a persistent level shift or structural capacity gap. Rather, it indicates that temporary periods of elevated average workload may warrant managerial attention. The daily results provide

more direct evidence of short-term capacity pressure, including the 42 working days on which demand exceeded nominal three-translator capacity.

The maximum observed daily demand would have required ten translators if the entire workload had to be processed within one working day. This extreme value likewise should not be interpreted as a permanent staffing target, because actual capacity requirements depend on due dates, acceptable backlog, prioritization, and the possibility of distributing work across several days. It instead illustrates the magnitude of short-term workload exposure under a same-day processing assumption.

The prediction intervals further illustrated this uncertainty. The 80% upper limit implied a requirement of six translators, whereas the 95% upper limit implied seven. Prediction intervals are not intended to identify a single deterministic staffing level; they describe a range of plausible demand outcomes and therefore provide a basis for contingency planning [38]. These results do not justify permanently increasing the team to six or seven translators. Rather, they quantify the temporary capacity that could be required under unusually high but plausible demand conditions.

The operational problem is consequently not limited to identifying one optimal workforce size. It involves configuring a stable core workforce for expected demand and a flexible capacity layer for exceptional workloads, productivity variation, and uncertain delivery requirements [4–6]. This distinction is especially important in knowledge-intensive services, where specialized capacity cannot always be expanded rapidly and where external support may involve coordination, quality, confidentiality, and availability constraints.

5.2. Theoretical and Practical Implications

5.2.1. Theoretical Implications

The first theoretical implication concerns integrating forecast evaluation and operational decision-making. Forecasting research commonly evaluates predictive performance through statistical error measures [20,22,23], whereas service-capacity research emphasizes resource allocation and demand–capacity alignment [3–6]. The present study connects these perspectives by expressing demand, forecast errors, productivity, and staffing requirements in compatible operational units.

This connection matters because statistical errors acquire managerial significance when translated into their operational consequences. In the present analysis, scale-dependent errors expressed in words were converted into translator-day equivalents of nominal productive capacity, allowing statistical forecast performance to be interpreted in the same unit used for workforce planning. These equivalents can subsequently be related to capacity shortage, unused capacity, and contingency requirements, although they should not be interpreted mechanically as discrete headcount changes. The study therefore supports a decision-oriented interpretation of model selection, in which the preferred model is evaluated not only by abstract predictive accuracy but also by the relevance of its errors to the decision being made [22–24].

This perspective also helps clarify why different accuracy measures should not be treated as interchangeable. MAE represents the typical absolute deviation in words, RMSE places greater weight on large errors, MASE assesses performance relative to a reference benchmark, and sMAPE provides proportional information. Their usefulness depends on the decision context. In workforce planning, scale-dependent errors are particularly valuable because they can be converted directly into working time and staffing requirements [22,23].

A second theoretical implication concerns the role of parsimony. The Mean model did not outperform the alternatives because it provided a complete representation of the demand-generating process. It performed better because the available data did not justify a

more responsive structure. Its superiority is therefore conditional on the short observation period, the irregular daily behavior of the series, and the absence of reliably identified seasonality or trend.

Parsimony should consequently not be interpreted as evidence that simple models are universally superior. Instead, it represents the avoidance of unsupported assumptions and unnecessary parameters. This interpretation is consistent with the principle that more elaborate models should demonstrate reliable out-of-sample improvements over transparent benchmarks before their additional complexity is considered justified [13,16–18].

A third contribution is the distinction between baseline and contingency capacity. Conventional demand–capacity calculations may suggest that one staffing value represents the operational optimum. The present results show that staffing derived from a point forecast represents only the core capacity required under expected conditions. Upper prediction limits, observed shortages, and productivity variation define a separate contingency-capacity problem.

This distinction extends the demand–capacity framework by connecting point forecasts with baseline staffing and probabilistic forecast ranges with flexible-capacity planning. It also supports the use of probabilistic forecasts rather than relying exclusively on point predictions, because decision-makers need to understand both the expected workload and the uncertainty surrounding it [38].

A fourth implication concerns measurement in knowledge-intensive services. Word count provided a common scale for heterogeneous documents and made it possible to connect workload with productive capacity. This is consistent with the need to define service-specific productivity indicators that are observable, standardized, and operationally interpretable [29]. However, the study also demonstrates that workload volume and workload effort are not equivalent.

The number of words captures the quantity of requested content but not the cognitive, technical, or temporal difficulty of completing the translation. Language direction, terminology, document format, revision requirements, urgency, and quality-control activities can cause equivalent word counts to require different amounts of productive time. Objective workload measurement is therefore necessary but insufficient for fully explaining effective capacity in professional services.

Finally, the study illustrates how a single-case design can contribute analytically when the organizational setting, data-generating process, assumptions, and limitations are made explicit [36]. The purpose is not statistical generalization to all translation services, but analytical examination of how benchmark forecasting and capacity conversion can be integrated within a specific operational context.

5.2.2. Practical Implications

The most immediate managerial implication is that the forecast-derived three-translator requirement should be interpreted as a baseline planning scenario rather than as evidence of consistently sufficient capacity. Under the nominal productivity assumption, this baseline covers the expected aggregate workload represented by the Mean forecast, but the historical exceedance frequency shows that it would remain exposed to periods of higher daily demand.

Permanent recruitment based on the maximum observed demand would probably generate substantial idle capacity because extreme workloads were infrequent and may be completed over several days. A more appropriate strategy is to maintain a core team capable of serving typical demand while establishing predetermined mechanisms for activating temporary capacity. These mechanisms may include a qualified pool of freelance

translators, cross-training, temporary internal reassignment, workload prioritization, and negotiated delivery dates [5,6].

The organization should define explicit activation thresholds rather than respond informally after overload has already occurred. For example, contingency support could be triggered when the scheduled workload exceeds available capacity for a defined number of consecutive days, when the accumulated backlog surpasses an agreed service threshold, or when forecast demand approaches the upper prediction limit. Because the same number of words has different capacity implications depending on the time available, activation rules should combine workload volume with required delivery dates.

The results also support daily rather than exclusively monthly monitoring. Monthly averages are useful for medium-term staffing discussions, but they conceal short-duration shortages, demand concentration, and extreme assignments. A management dashboard should therefore report current demand, forecast demand, available capacity, utilization, accumulated backlog, forecast error, due dates, and the number of days on which nominal capacity was exceeded.

Data-quality controls should accompany this monitoring process. The reliability of future forecasts depends on the accuracy, completeness, consistency, and temporal comparability of the underlying operational records [37]. Standardized recording should therefore distinguish actual zero-demand days from missing records and should document language direction, document type, urgency, deadline, and completed workload wherever possible.

The sensitivity analysis also shows that the baseline staffing estimate should not be interpreted independently of the productivity assumption. At 1600 words per translator per day, the point forecast requires four translators rather than three, whereas the nominal and favorable scenarios both round to three. Managers should therefore treat 2000 words per day as a provisional planning parameter and progressively replace it with empirically estimated productivity and availability distributions differentiated by language direction, document type, complexity, revision requirements, urgency, and non-translation time.

Finally, the data-to-decision procedure provides a practical governance mechanism. Standardized data capture, validation, forecasting, capacity conversion, action, and monitoring create a continuous organizational learning cycle rather than a one-time analytical exercise [9–11]. As additional observations become available, managers can reassess benchmark performance, identify emerging temporal patterns, update productivity assumptions, and determine whether the Mean model remains suitable.

5.3. Limitations and Future Research

This study has several limitations that define the scope of its conclusions.

First, the dataset covered only six months. Although it contained 122 daily observations, it did not include repeated annual periods and therefore could not support reliable conclusions about annual seasonality, long-term cycles, or structural changes. The July–December forecasts should consequently be interpreted as planning scenarios based on a stable-level benchmark rather than as externally validated long-horizon predictions. The retained constant-level specification was also not formally tested against possible level shifts or systematic day-of-week effects. The observed monthly differences therefore cannot be interpreted as evidence of a structural change in the demand process. Future analysis could examine these possibilities by estimating specifications with change-point or intervention terms for potential level shifts and weekday indicators for recurring within-week effects, and by comparing their out-of-sample performance with the current benchmarks under the same rolling-origin evaluation framework.

Future research should continue daily data collection for at least one complete annual cycle and preferably for several years. A longer series would permit the evaluation of weekly, monthly, and annual patterns, calendar effects, institutional deadlines, changes in demand level, and possible structural breaks. It would also enable more reliable comparison of seasonal and nonseasonal models [15,34].

Second, the study examined a single translation team in one organizational context. Its demand process, workflow, staffing structure, service requirements, and productivity conditions may differ from those of commercial translation agencies, universities, public institutions, or multilingual service centers. In accordance with the logic of single-case research, the findings support analytical rather than statistical generalization [36], Ch. 2. Replication across teams and organizations is required to evaluate the transferability of the procedure under different workload distributions and service-level conditions.

Third, demand was represented exclusively through the number of words requested. This measure improved comparability across document formats but did not capture document complexity, language direction, terminology, formatting, quality assurance, revision intensity, urgency, or delivery time. Two assignments with the same number of words may require substantially different levels of professional effort.

Future studies should develop workload-equivalent measures that combine word count with complexity or processing-time weights. Multivariate models could subsequently include predictors such as document type, requesting unit, language pair, urgency, expected deadline, academic calendar, and known organizational events. These variables should be included only when they are available at the forecast origin to avoid data leakage [20].

Fourth, the individual capacity assumption of 2000 words per day was derived from a one-week observation and professional consensus rather than from a longitudinal productivity study. Although the sensitivity analysis examined $\pm 20\%$ scenarios, these scenarios do not replace empirical estimation. Future work should measure individual and team productivity over longer periods and distinguish direct translation time from meetings, coordination, revision, interruptions, leave, and other responsibilities. In addition, the analysis treated individual productivity and workforce availability separately only in conceptual terms. It did not model day-specific reductions in available headcount caused by annual leave, illness, training, meetings, or other organizational duties. Future studies should therefore estimate both productive output conditional on translator availability and the probability distribution of available workforce, allowing effective team capacity to vary across working days rather than treating it as a fixed nominal parameter.

Capacity could then be modeled as a probability distribution rather than as a fixed constant. Such an approach would permit the uncertainty in both demand and productivity to be incorporated into staffing decisions, extending the current use of prediction intervals for demand alone [38].

Fifth, the available records contained only an aggregate demand series and did not provide separate workloads for English and Spanish. Accordingly, all staffing estimates in the present study refer to the combined translation workload, and no language-specific headcount requirement can be inferred from the current analysis. Future data collection should identify source language, target language, assigned translator, processing time, completion date, and revision requirements. Separate forecasts could then determine whether demand patterns and capacity requirements differ across language directions.

Sixth, the rolling-origin evaluation was intentionally restricted to the Mean, Naive, and Drift benchmarks to establish transparent reference performance. This was a methodological choice rather than a constraint imposed by the number of observations, since standard alternatives such as optimized simple exponential smoothing, Holt/ETS, ARIMA, Theta, and forecast combinations can also be estimated for series of this length [12,13].

The principal limitation is instead the six-month temporal coverage, which restricts the reliable validation of recurring longer-term patterns. Future research should compare these alternative specifications with the present benchmarks using the same rolling-origin framework and determine whether their additional structure yields consistent out-of-sample improvements [20,21]. As longer histories become available, such comparisons could also evaluate seasonal, calendar, level-shift, and other temporal effects more reliably.

Machine-learning or neural approaches should be considered only when the enlarged dataset and predictor structure can support their estimation. Their usefulness should be demonstrated through temporally valid out-of-sample improvements rather than inferred from model complexity [16–18,20,21].

Seventh, the study converted daily demand directly into same-day workforce requirements. Translation assignments may, however, be completed over different time windows, and some backlog may remain acceptable within negotiated deadlines. A more complete operational model should therefore incorporate due dates, work in process, queue length, service priorities, translator specialization, outsourcing lead times, and asymmetric costs of undercapacity and overcapacity.

Future research could complement demand forecasting with queueing, simulation, or optimization approaches to evaluate alternative staffing and flexible-capacity policies under explicit service-level conditions [5–7]. Such approaches could distinguish the capacity required for same-day processing from that required when backlog or more flexible delivery commitments are permitted.

Finally, the study did not measure the organizational outcomes produced by implementing the proposed data-to-decision procedure. Future intervention-based research should evaluate whether continued use of the system reduces backlog, overtime, delayed deliveries, emergency outsourcing, or perceived workload pressure. Such evidence would extend the assessment beyond predictive accuracy and determine the actual operational value of the forecasting and monitoring process.

6. Conclusions

This study evaluated how parsimonious benchmark forecasting models can support workforce capacity planning in a translation service with limited historical data. Using 122 daily demand observations, the analysis compared the Mean, Naive, and Drift methods through expanding-window rolling-origin evaluation and translated forecast demand into staffing requirements.

The Mean method consistently produced the lowest MAE, RMSE, MASE, and sMAPE across one-, five-, and twenty-working-day horizons. Its superior performance among the three benchmarks evaluated indicates that a stable-level representation was more appropriate than persistence or linear-drift representations for the observed series. This conclusion is conditional on the short temporal observation window and highly variable daily demand and should be reassessed as additional data become available and alternative forecasting specifications are evaluated.

The final point forecast was 5472.12 words per working day. At the nominal productivity assumption of 2000 words per translator per day, this workload corresponded to 2.74 translator-equivalents and therefore supported a baseline requirement of three translators for the aggregate workload. However, observed demand exceeded the forecast-derived three-translator baseline capacity on 34.43% of working days, and the upper prediction limits implied substantially higher temporary requirements. The baseline estimate was also sensitive to the productivity assumption: under the conservative scenario of 1600 words per translator per day, the point-forecast requirement increased to four translators. The three-translator baseline should therefore be interpreted as sufficient for expected aggregate

demand under the nominal productivity assumption, but not as a fixed staffing prescription or as sufficient to cover all short-term demand peaks.

The central managerial conclusion is that baseline capacity and peak-capacity management should be treated separately. A forecast-derived baseline of three translator-equivalents can address the typical aggregate workload under the nominal assumptions, while flexible arrangements may be required to absorb unusually large requests, productivity variation, and periods of accumulated demand. Such arrangements may include freelance support, workload prioritization, negotiated deadlines, cross-training, and continuous monitoring.

The study also demonstrates that forecast evaluation becomes more meaningful when predictions are expressed in the operational units used for decision-making. By linking daily demand, forecast error, translator productivity, and workforce requirements, the proposed data-to-decision procedure provides a transparent foundation for capacity planning in a knowledge-intensive service. Its principal value lies not in identifying a universally superior model or staffing level, but in establishing a reproducible process that can be updated as new information becomes available.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/forecast8050085/s1, Table S1: Daily translation-demand observations from January to June 2025. The table contains the 122 working-day observations used in the descriptive analysis, rolling-origin forecast evaluation, final demand forecasting, and workforce-capacity calculations, including the sequential observation number, date, and total daily demand measured in words.

Author Contributions: Conceptualization, L.H.-M.; methodology, D.R.T.-V., R.M.-S., L.H.-M. and R.P.-C.; software, G.G.-V. and R.P.-C.; validation, D.R.T.-V., R.M.-S. and A.S.-R.; formal analysis, L.H.-M., R.P.-C. and G.G.-V.; investigation, A.S.-R., L.H.-M., R.P.-C., D.R.T.-V. and R.M.-S.; resources, G.G.-V. and A.S.-R.; data curation, R.M.-S., R.P.-C., L.H.-M. and G.G.-V.; writing—original draft preparation, L.H.-M.; writing—review and editing, A.S.-R.; visualization, D.R.T.-V. and R.M.-S.; supervision, G.G.-V.; project administration, R.P.-C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data supporting the findings of this study are provided in the Supplementary Materials as Table S1, which contains the 122 daily translation-demand observations used in the descriptive analysis, rolling-origin forecast evaluation, final forecasting, and workforce-capacity calculations. Additional information regarding data preparation and analytical procedures may be obtained from the corresponding author upon reasonable request.

Acknowledgments: The authors thank the translation-team coordinator and participating professionals for facilitating access to the operational records and providing contextual information regarding workload allocation and productive capacity. During the preparation of this manuscript, the authors used ChatGPT (OpenAI, GPT-5.6 Thinking) to support translation, language editing, wording refinement, manuscript restructuring, methodological clarification, and verification of statistical reporting. The authors reviewed and independently verified all outputs, calculations, interpretations, and references and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Chase, R.B.; Apte, U.M. A history of research in service operations: What's the big idea? *J. Oper. Manag.* **2007**, *25*, 375–386. [CrossRef]
2. Sampson, S.E.; Froehle, C.M. Foundations and implications of a proposed unified services theory. *Prod. Oper. Manag.* **2006**, *15*, 329–343. [CrossRef]

3. Klassen, K.J.; Rohleder, T.R. Demand and capacity management decisions in services: How they impact on one another. *Int. J. Oper. Prod. Manag.* **2002**, *22*, 527–548. [CrossRef]
4. Campbell, D.; Frei, F. Market heterogeneity and local capacity decisions in services. *Manuf. Serv. Oper. Manag.* **2011**, *13*, 2–19. [CrossRef]
5. Li, G.; Field, J.M.; Jiang, H.; He, T.; Pang, Y. Decision models for workforce and technology planning in services. *Serv. Sci.* **2015**, *7*, 29–47. [CrossRef]
6. Yuan, Y. Managing flexible capacity in service systems with worker shortages. *Manuf. Serv. Oper. Manag.* **2025**, *27*, 808–824. [CrossRef]
7. Hopp, W.J.; Spearman, M.L. *Factory Physics*, 3rd ed.; Waveland Press: Long Grove, IL, USA, 2011.
8. Slack, N.; Chambers, S.; Johnston, R. *Operations Management*, 6th ed.; Pearson Education: Harlow, UK, 2009.
9. Brynjolfsson, E.; McElheran, K. The rapid adoption of data-driven decision-making. *Am. Econ. Rev.* **2016**, *106*, 133–139. [CrossRef]
10. Provost, F.; Fawcett, T. *Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking*; O'Reilly Media: Sebastopol, CA, USA, 2013.
11. Wamba, S.F.; Gunasekaran, A.; Akter, S.; Ren, S.J.; Dubey, R.; Childe, S.J. Big data analytics and firm performance: Effects of dynamic capabilities. *J. Bus. Res.* **2017**, *70*, 356–365. [CrossRef]
12. Petropoulos, F.; Apiletti, D.; Assimakopoulos, V.; Babai, M.Z.; Barrow, D.K.; Ben Taieb, S.; Bergmeir, C.; Bessa, R.J.; Bijak, J.; Boylan, J.E.; et al. Forecasting: Theory and practice. *Int. J. Forecast.* **2022**, *38*, 705–871. [CrossRef]
13. Hyndman, R.J.; Athanasopoulos, G. *Forecasting: Principles and Practice*, 3rd ed.; OTexts: Melbourne, Australia, 2021.
14. Shmueli, G. To explain or to predict? *Stat. Sci.* **2010**, *25*, 289–310. [CrossRef]
15. Box, G.E.P.; Jenkins, G.M.; Reinsel, G.C.; Ljung, G.M. *Time Series Analysis: Forecasting and Control*, 5th ed.; John Wiley & Sons: Hoboken, NJ, USA, 2015.
16. Makridakis, S.; Spiliotis, E.; Assimakopoulos, V. Statistical and machine learning forecasting methods: Concerns and ways forward. *PLoS ONE* **2018**, *13*, e0194889. [CrossRef] [PubMed]
17. Makridakis, S.; Spiliotis, E.; Assimakopoulos, V. The M4 Competition: 100,000 time series and 61 forecasting methods. *Int. J. Forecast.* **2020**, *36*, 54–74. [CrossRef]
18. Petropoulos, F.; Grushka-Cockayne, Y.; Siemsen, E.; Spiliotis, E. Wielding Occam's razor: Fast and frugal retail forecasting. *J. Oper. Res. Soc.* **2025**, *76*, 1564–1583. [CrossRef]
19. Svetunkov, I.; Petropoulos, F. Old dog, new tricks: A modelling view of simple moving averages. *Int. J. Prod. Res.* **2018**, *56*, 6034–6047. [CrossRef]
20. Hewamalage, H.; Ackermann, K.; Bergmeir, C. Forecast evaluation for data scientists: Common pitfalls and best practices. *Data Min. Knowl. Discov.* **2023**, *37*, 788–832. [CrossRef] [PubMed]
21. Tashman, L.J. Out-of-sample tests of forecasting accuracy: An analysis and review. *Int. J. Forecast.* **2000**, *16*, 437–450. [CrossRef]
22. Hyndman, R.J.; Koehler, A.B. Another look at measures of forecast accuracy. *Int. J. Forecast.* **2006**, *22*, 679–688. [CrossRef]
23. Koutsandreas, D.; Spiliotis, E.; Petropoulos, F.; Assimakopoulos, V. On the selection of forecasting accuracy measures. *J. Oper. Res. Soc.* **2022**, *73*, 937–954. [CrossRef]
24. den Boer, A.V.; Sierag, D.D. Decision-based model selection. *Eur. J. Oper. Res.* **2021**, *290*, 671–686. [CrossRef]
25. Fitzsimmons, J.A.; Fitzsimmons, M.J. *Service Management: Operations, Strategy, Information Technology*, 8th ed.; McGraw-Hill Education: New York, NY, USA, 2014.
26. Heizer, J.; Render, B.; Munson, C. *Operations Management: Sustainability and Supply Chain Management*, 13th ed.; Pearson: Harlow, UK, 2020.
27. Shmueli, G.; Koppius, O.R. Predictive analytics in information systems research. *MIS Q.* **2011**, *35*, 553–572. [CrossRef]
28. Pérez-Campdesuñer, R.; Sánchez-Rodríguez, A.; García-Vidal, G.; Martínez-Vivar, R.; De Miguel-Guzmán, M. Artificial Intelligence in Ecuadorian SMEs: Drivers and Obstacles to Adoption. *Information* **2025**, *16*, 443. [CrossRef]
29. Johnston, R.; Clark, G. *Service Operations Management: Improving Service Delivery*, 3rd ed.; Financial Times/Prentice Hall: Harlow, UK, 2008.
30. Gaither, N.; Frazier, G. *Operations Management*, 9th ed.; South-Western College Publishing: Cincinnati, OH, USA, 2001.
31. Lovelock, C.H.; Wright, L. *Principles of Service Marketing and Management*, 2nd ed.; Prentice Hall: Upper Saddle River, NJ, USA, 2001.
32. Moreira, D.A. *Administração da Produção e Operações*; Pioneira Thomson Learning: São Paulo, Brazil, 2004.
33. Sánchez-Rodríguez, A.; Martínez-Vivar, R.; Moreno-Lázaro, J. Gestión por competencias laborales en el contexto del proceso de cambios políticos y económicos en Cuba. *Innovar* **2017**, *27*, 169–184. [CrossRef]
34. Morettin, P.A.; Toloi, C.M.C. *Análise de Séries Temporais: Modelos Lineares Univariados*, 3rd ed.; Blucher: São Paulo, Brazil, 2018; Volume 1.
35. Kyriazi, F.; Thomakos, D.D. Benchmarks in univariate time-series forecasting: A historical and methodological review. *IMA J. Manag. Math.* **2026**, *37*, 327–374. [CrossRef]

36. Yin, R.K. *Case Study Research and Applications: Design and Methods*, 6th ed.; SAGE Publications: Thousand Oaks, CA, USA, 2018.
37. Wang, R.Y.; Strong, D.M. Beyond accuracy: What data quality means to data consumers. *J. Manag. Inf. Syst.* **1996**, *12*, 5–33. [CrossRef]
38. Gneiting, T.; Katzfuss, M. Probabilistic forecasting. *Annu. Rev. Stat. Appl.* **2014**, *1*, 125–151. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.