forecasting

MDPI

Article

Pooling Beats Structure on Short Annual Panels: Parameter Sharing and the Forecast Accuracy of Evolutionary Models for Compositional Time Series

Aras Yolusever

Department of Economics, Faculty of Economics and Administrative Sciences, Istanbul Kültür University, 34303 Istanbul, Türkiye; a.yolusever@iku.edu.tr

Highlights

What are the main findings?

- Sharing parameters across countries reduces forecast error by 15.3% to 27.8% across three Eurostat panels, an effect larger than the gap between the best structural model and the best statistical benchmark.
- Replicator–mutator structure does not by itself improve accuracy; it ties the random walk with drift, which it nests as a special case, to educational attainment and loses to a no-change forecast on municipal waste routes. Pooled kernel and neural benchmarks finish mid-field on every panel.

What are the implications of the main findings?

- On short annual panels of official statistics, estimating evolutionary models country by country is the single choice most damaging to out-of-sample accuracy; a validation-selected parameter-sharing regime corrects it, and a simulation grid shows that the gain hinges on limited cross-country parameter heterogeneity.
- The shrinkage weight and the mutation rate are tuning parameters rather than structural constants and should be selected by walk-forward optimisation inside the training window; the estimated congestion parameters keep their economic reading, increasing returns in attainment and congestion in waste routes.

Abstract

Many economic time series are compositions: educational attainment shares, waste treatment routes, energy mixes, and sectoral employment. Forecasters usually move such series to log-ratio coordinates and apply generic methods that impose no economic restriction on how the parts move. Evolutionary game theory supplies one: replicator dynamics make the growth rate of a share proportional to its payoff advantage, which is the logic of imitation, diffusion and congestion that economics itself uses to explain why shares move. We turn that restriction into a forecast function with $K + 1$ parameters for a K-part composition; estimate it under three parameter-sharing regimes, namely country-specific, fully pooled and validation-shrunk; and race it against classical log-ratio benchmarks, a non-evolutionary Dirichlet comparator and two pooled machine-learning benchmarks in an expanding-window, rolling-origin design on three Eurostat panels. The evolutionary restriction does not buy accuracy: its best specification ties the random walk with drift on educational attainment and loses to persistence on municipal waste routes. What moves accuracy is the sharing regime. Moving from country-specific estimation to the best sharing regime cuts the mean absolute scaled error by 27.8% on attainment and by 15.3% to 19.8% on the waste panels, more than the gap between the best structural and the best statistical

Check for updates

Academic Editor: Dimitrios D. Thomakos

Received: 7 August 2026
Revised: 14 September 2026
Accepted: 15 September 2026
Published: 18 September 2026

Copyright: © 2026 by the author. 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.

model, and the same ordering reappears in the Dirichlet family. The estimated congestion parameters carry the economics the restriction was built for, increasing returns in attainment and congestion in waste routes, while a simulation grid shows the sharing gain reversing once cross-country heterogeneity is appreciable. We close with a walk-forward procedure for tuning the shrinkage weight and the mutation rate.

Keywords: economic theory; compositional time series; replicator–mutator dynamics; forecast evaluation; parameter pooling; shrinkage; walk-forward validation; model confidence set; evolutionary game theory

1. Introduction

A large share of the quantities that economists forecast are not levels but shares. The educational composition of the working-age population, the split of municipal waste across treatment routes, the fuel mix of an electricity system, the distribution of employment across contract types: in each case the object of interest is a vector of non-negative parts that sums to one. Such vectors live on the unit simplex, and forecasting them with methods designed for unconstrained series risks predictions that are negative, that fail to add up, or that drift outside the feasible region at long horizons.

The standard response, following Aitchison [1,2], is to move to log-ratio coordinates, forecast there with a generic method, and map the result back. This solves the coherence problem completely and cheaply. What it does not do is impose any economic content. A vector autoregression in log-ratio coordinates treats the composition as an arbitrary multivariate process; nothing in its structure says that a part should grow when the activity it represents is doing comparatively well.

Economics has a natural candidate for exactly that restriction. Replicator dynamics, introduced by Taylor and Jonker [3] and developed into the standard apparatus of evolutionary game theory by Hofbauer and Sigmund [4], Weibull [5] and Sandholm [6], make the growth rate of a share proportional to the gap between its payoff and the population average payoff. Its behavioural microfoundations are well understood: it arises from imitation of successful others under bounded rationality [7] and from payoff-monotone revision protocols more generally [6]. Adding mutation, in the sense of Nowak, Komarova and Niyogi [8] and Komarova [9], introduces a small flow toward strategies that are not currently doing well, which can be read as experimentation, entry, or simply the failure of imitation to be perfect. Section 2 shows that this restriction is not an import from biology: the logistic diffusion curve, the substitution model of technological change and the congestion game are all special cases of it, and economics has been using them to explain moving shares for seventy years.

Why this matters is a question about practice rather than about theory. Evolutionary and replicator specifications are now routinely fitted to shares taken from official statistics, and the standard reporting pattern in that applied literature is an in-sample fit, estimated unit by unit, with the model judged by how closely the fitted path tracks the observed one. Nothing in that pattern tells a user whether the restriction helps on data that have not yet been seen, which is the only sense in which a forecasting model earns its keep, and nothing tells the user whether the effort spent on specifying payoffs would be better spent elsewhere. Two questions are therefore open, and the paper is organised around them. First, does an economic restriction on how shares move buy out-of-sample accuracy against the generic transformations that applied compositional forecasting actually uses? Second, if it does not, where does accuracy in this data environment come from instead?

https://doi.org/10.3390/forecast8050089

The answer to the second question is what makes the answer to the first useful, because it identifies the design margin a practitioner can still move.

We treat the question as a forecasting question, with a rolling-origin design, a panel of benchmarks, scale-free losses and formal tests of equal predictive ability. The answer is not the one the framing invites. The restriction does not pay for itself in accuracy: its best specification ties the random walk with drift on educational attainment, which turns out to be the special case of the evolutionary map with no frequency dependence and no mutation, and it loses to a no-change forecast on municipal waste routes. What pays is parameter sharing. Across three Eurostat panels, the spread within the structural family generated purely by how parameters are shared across countries is 18.1% to 38.5% in mean absolute scaled error, whereas the gap between the best structural specification and the best benchmark is 1.3% on one panel and 11.5% to 13.1% on the others. Moving from country-specific estimation to the best sharing regime reduces error by 27.8% on attainment and by 15.3% and 19.8% on the two waste panels, and the same ordering reappears when a non-evolutionary Dirichlet transition model is estimated under identical regimes.

It is worth saying plainly what is new here and what is not. Replicator–mutator dynamics, parameter pooling, shrinkage and the bias-variance trade-off that governs them are all established, and this paper introduces none of them. What it contributes is the controlled comparison itself: an economic restriction, a non-evolutionary simplex-native likelihood model and a set of classical log-ratio benchmarks are estimated under identical pooling regimes, at identical rolling origins, under identical losses and with identical treatment of zero parts, so that the model-class margin and the parameter-sharing margin are measured on one scale and can be read off one table. That comparison does not appear to have been made for compositional time series, and it is what allows the ordering claim to be stated at all. The claim is bounded: a simulation grid shows the ordering reversing once cross-country parameter heterogeneity is appreciable, so the result is a statement about data environments in which estimation uncertainty dominates heterogeneity, which is where short annual official panels appear to live.

The paper is organised so that the economics comes first and the technicalities last. Section 2 explains why an evolutionary restriction makes sense for compositional forecasts and gives three theoretical examples. Section 3 states the forecast function in a single equation, places the benchmarks in the same notation, defines the sharing regimes and describes the data and the evaluation protocol. Section 4 reports one coherent set of results and reads it through economic intuition. Section 5 places the result at the intersection of the economics and the forecasting literature and states its scope. Section 6 proposes a walk-forward procedure for tuning the parameters that the evidence identifies as tuning devices. Section 7 concludes. Appendix A collects the simplex geometry, the stability analysis, the estimation and benchmark details, the inferential procedures and the Monte Carlo evidence; the Supplementary Materials hold the horizon-by-horizon accuracy tables, the win-rate analysis, the estimated parameters and the robustness checks.

2. Why an Evolutionary Restriction Makes Economic Sense

2.1. What the Restriction Encodes

Let $x_t = (x_{1,t}, \ldots, x_{K,t})$ be the shares of K alternatives at date t. The replicator restriction says that the share of alternative k grows between t and $t+1$ in proportion to how much better it pays than the population average, and the mutator term says that a small fraction of every share is reallocated uniformly regardless of payoff. Three economic readings of this restriction are standard, and each of them is a reason to expect it to describe the compositions that official statistics report.

The first reading is imitation. Schlag [7] showed that a population of boundedly rational agents who sample one other agent and switch to that agent's alternative with a probability proportional to the payoff difference generates replicator dynamics in the large-population limit; Fudenberg and Levine [10] and Young [11] develop the same idea into a theory of learning in populations, and Sandholm [6] shows that any payoff-monotone revision protocol produces dynamics of the replicator class. When households choose between attainment levels, or firms and municipalities between disposal routes, the relevant decision-makers observe the outcomes of others and move toward what works, and imitation of the successful is the mechanism through which aggregate shares move.

The second reading is selection among routines. Nelson and Winter [12] built evolutionary economics on the observation that the market share of a technique or a firm grows with its relative profitability, so that the composition of an industry is the outcome of a selection process rather than of a representative optimiser. The replicator equation is the reduced form of that process: the growth rate of a share equals its payoff advantage, with no agent required to know the payoffs of alternatives it does not currently use.

The third reading is diffusion. With two alternatives and a constant payoff gap δ, the replicator restriction implies that the log-odds of the two shares grows linearly, $\ln(x_{1,t+1}/x_{2,t+1}) = \ln(x_{1,t}/x_{2,t}) + \delta$, which is the logistic curve. Griliches [13] fitted exactly this curve to the diffusion of hybrid corn across American states and read its slope as the profitability of adoption; Fisher and Pry [14] proposed the same linear log-odds law as a general model of technological substitution and used it to forecast the replacement of one technology by another. Both are replicator dynamics under another name, and both have been forecasting devices from the start. The K-part map used in this paper is their multi-alternative generalisation with two additions: a congestion term that lets the payoff of an alternative depend on how many already use it, and a mutation term that keeps every alternative alive.

The congestion term is where the economics of the composition is encoded. Payoffs are linear in the composition with an alternative-specific intercept and a common congestion parameter, $\pi_k(x) = \alpha_k - \gamma x_k$. A positive γ means that an alternative becomes less attractive the more of the population uses it, which is congestion, crowding, or diminishing returns to a route, and it makes an interior rest point of the map locally stable, as Appendix A.2 shows in closed form. A negative γ means increasing returns, so that adoption is self-reinforcing: this is the lock-in mechanism of Arthur [15], under which historical accident can select among several stable long-run compositions, and it renders the interior rest point unstable and moves the attractors toward the faces of the simplex. The sign of a single parameter therefore separates two families of economic stories, and Section 4 shows that the data sort the panels between them in the way the stories predict.

2.2. Three Theoretical Examples

The first example is educational attainment. Consider the working-age population sorted into low, medium and tertiary attainment. If the payoff to a level rose only with its scarcity, the composition would settle at an interior rest point, and the share of the tertiary level would stop growing once its wage premium was competed away. The theory of directed technical change says otherwise: Acemoglu [16] shows that a larger supply of skilled workers enlarges the market for skill-complementary technologies and can raise rather than lower the skill premium, so that the payoff to an attainment level increases with the share that already holds it. Peer effects in schooling and the co-evolution of attainment with the occupational structure that rewards it point the same way. In the notation of the map, this is $\gamma < 0$: the attainment composition should trend monotonically toward the tertiary corner along an S-shaped path, and the estimated congestion parameter on the

attainment panel should be negative. Section 4.2 reports that it is negative for 23 of 33 countries.

The second example is municipal waste. Treated waste is split across recycling, incineration and landfill. Each route has a marginal cost that rises with the load it carries: recycling capacity is limited by sorting and reprocessing plants, incineration by furnace capacity, landfill by remaining volume and by gate fees and taxes that increase with use. A municipality that observes the marginal cost of each route directs waste to the cheapest one at the margin, so the share of a route grows when its cost is below the average, which is the replicator restriction with $\gamma > 0$. The rest point of the map is the route mix at which marginal costs are equalised across the routes in use, the condition familiar from congestion games and traffic assignment, and policy shocks such as a landfill tax enter as shifts in the intercepts α_k. The estimated congestion parameter on the waste panels should be positive, and Section 4.2 reports medians of 0.109 and 0.244.

The third example is the fuel mix of an electricity system, or any other technology substitution. Fisher and Pry [14] showed that once a new technology has captured a few per cent of a market, the log-odds of its share grows at a steady rate until it dominates; the replicator map with a constant payoff gap reproduces this exactly, and the congestion term adds the one feature the logistic lacks: a mechanism by which the substitution can stop short of complete displacement when the incumbent retains a cost advantage at low load. The same reasoning applies to employment shares across contract types or sectors, where the payoff to a form of employment depends on relative wages and on how crowded the form already is. None of these compositions is in the data used below, but they are the compositions to which the applied literature fits replicator specifications, and the point of the examples is that the restriction is not an analogy but the reduced form of the economic stories that are told about these series anyway.

2.3. Parameter Sharing as an Economic Prior

The restriction has a second economic implication that the applied literature usually ignores. The map has $K - 1$ intercepts, one congestion parameter and one mutation rate, hence $K + 1$ free parameters, against $m(m + 1)$ for the conditional mean of a first-order vector autoregression in $m = K - 1$ log-ratio coordinates. The promise of the restriction as a forecasting device is therefore variance reduction: it buys fewer parameters with economic structure. Whether the purchase pays is an empirical question, and it is the first question of the paper. But the same logic applies to how the parameters are estimated. The intercepts measure the relative attractiveness of alternatives, which plausibly differs across countries with different institutions and prices; the congestion parameter and the mutation rate measure the technology of adjustment, how strongly payoffs respond to crowding and how imperfectly agents imitate, which plausibly does not. Countries drawing on a common statistical system, a common regulatory space and, for the waste panels, a common set of directives are therefore natural candidates for sharing the parameters of adjustment, and the choice between estimating the map country by country or pooling it across countries is itself an economic prior about heterogeneity. Griliches [13] already faced this choice when he fitted separate logistic curves state by state and then explained the differences in their slopes by profitability; Garcia-Ferrer, Highfield, Palm and Zellner [17] and Zellner and Hong [18] showed that shrinking country-specific coefficients toward a pooled mean improves international forecasts. Section 3.3 turns this prior into three sharing regimes, and Section 4 shows that the choice among them moves accuracy more than the restriction itself.

3. Materials and Methods

3.1. The Forecast Function

Let $x_{i,t}$ be the composition of unit i at date t, with K strictly positive parts summing to one. The structural forecast at horizon h from origin t is the h-fold iterate of a single one-step map,

$$\hat{x}_{i,t+h|t} = F^h(x_{i,t};\theta), \qquad F_k(x;\theta) = (1-\mu)\,\frac{x_k \exp(\alpha_k - \gamma x_k)}{\sum_{j=1}^{K} x_j \exp(\alpha_j - \gamma x_j)} + \frac{\mu}{K}, \qquad \alpha_K = 0, \quad (1)$$

with parameter vector $\theta = (\alpha_1, \ldots, \alpha_{K-1}, \gamma, \mu)$. Every symbol in Equation (1) has a role. The intercepts α_k are the relative attractiveness of alternative k against the reference alternative K; the normalisation $\alpha_K = 0$ is required because only payoff differences matter. The congestion parameter γ makes the attractiveness of an alternative depend on its own current share, with the sign interpretation of Section 2.1. The selection step reweights the current shares by exponential fitness, which keeps every iterate strictly inside the simplex and is numerically stable when the map is iterated five steps ahead; a selection intensity multiplying the payoffs is not separately identified from the scale of (α, γ) and is normalised to one. The mutation rate μ mixes the selected composition toward the uniform composition $1/K$, so that no alternative ever dies out and, in forecasting terms, so that multi-step forecasts are shrunk toward the centre of the simplex. Figure 1 sets out one application of the map in schematic form, from the observed composition through payoffs, selection and exploration to the next composition, together with the recursion that produces multi-step forecasts. Every intermediate quantity is a composition, so the forecasts cohere at every horizon without renormalisation.

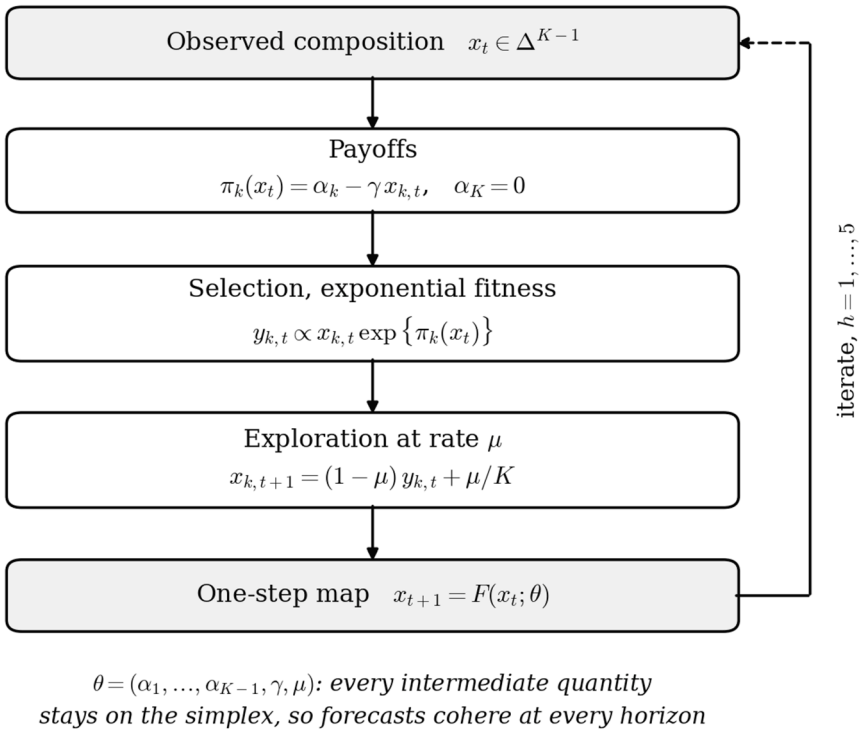


Figure 1. One step of the structural forecast function of Equation (1). The composition observed at the origin enters the payoff function; selection reweights the parts by exponential fitness; exploration mixes the result toward the uniform composition at the mutation rate; the resulting composition is fed back into the map, which is how forecasts at horizons one through five are constructed.

The same map is easier to compare with the benchmarks in log-ratio coordinates. With $\mu = 0$, taking the log-ratio of any part to the reference part on both sides of Equation (1) gives

$$\ln \frac{\hat{x}_{k,t+1|t}}{\hat{x}_{K,t+1|t}} = \ln \frac{x_{k,t}}{x_{K,t}} + \alpha_k - \gamma(x_{k,t} - x_{K,t}), \qquad k = 1, \ldots, K-1. \tag{2}$$

Equation (2) says that the structural model is a random walk with drift in additive log-ratio coordinates whose drift is state-dependent: the constant part α_k is the drift of the classical benchmark, the term $-\gamma(x_{k,t} - x_{K,t})$ is the frequency-dependent correction that the economics adds, and the mutation term of Equation (1) is a shrinkage of the whole vector toward the barycentre. The random walk with drift is therefore the special case $\gamma = 0$, $\mu = 0$, and the no-change forecast is the special case $\alpha = 0$, $\gamma = 0$, $\mu = 0$. This nesting is what makes the results of Section 4 interpretable: the evolutionary restriction can only beat these two benchmarks if the curvature γ and the smoothing μ are worth their estimation cost on the data at hand.

The parameters are estimated by conditional least squares in the geometry of the simplex. Writing $d_A(u, v) = \| \operatorname{clr}(u) - \operatorname{clr}(v) \|_2$ for the Aitchison distance, the Euclidean distance between centred log-ratio images, the estimator at origin t_0 is

$$\hat{\theta} = \operatorname*{argmin}_{\theta} \sum_{i \in \mathcal{S}} \sum_{t < t_0} d_A(x_{i,t+1}, F(x_{i,t}; \theta))^2, \tag{3}$$

where $\mathcal{S}$ is the set of units whose training data enter the objective, and it is through $\mathcal{S}$ that the sharing regime of Section 3.3 enters. The mutation rate is estimated on the logistic scale $\widetilde{\mu} = \ln(\mu/(1-\mu))$ so that the parameter vector is unconstrained, and the optimiser, the starting values and the treatment of zero parts are described in Appendix A.3. Multi-step forecasts are produced by iterating the estimated map from the last observed composition, which yields a coherent path on the simplex at every horizon.

3.2. The Benchmarks in the Same Notation

Every benchmark has the same structure as Equation (1): a one-step map, iterated h times, applied to the composition or to its log-ratio image. Writing $z_t = \operatorname{alr}(x_t)$ for the additive log-ratio coordinates, $z_k = \ln(x_k / x_K)$, the classical benchmarks forecast

$$\hat{x}_{t+h|t} = \operatorname{alr}^{-1}\Big(G^h(z_t; \phi)\Big), \tag{4}$$

where in Equation (4), G is the benchmark's one-step map in log-ratio coordinates and ϕ its parameter vector, and the inverse transform returns the forecast to the simplex, so that every forecast in the paper, structural or statistical, lies on the simplex by construction. Table 1 lists the one-step map and the parameter vector of each competitor and counts the free parameters at $K = 3$ and $K = 4$. The no-change benchmark repeats the last observed composition. The random walk with drift extrapolates the average log-ratio change over the training window. Univariate ARIMA and exponential smoothing models are fitted per coordinate with orders and trend components selected by corrected Akaike information [19]. A vector autoregression is fitted to the full log-ratio vector with lag order chosen by Akaike information. The dynamic Dirichlet comparator is a non-evolutionary transition model specified directly on the simplex, in the tradition of Grunwald, Raftery and Guttorp [20] and of the Dirichlet autoregressive models developed from it [21]: the next composition is Dirichlet distributed about a mean that follows a first-order recursion linear in the Aitchison geometry, with $K - 1$ location parameters, a scalar persistence parameter and a precision, hence exactly $K + 1$ free parameters, so that the comparison between the two dynamic families is not confounded with a difference in parameter counts; its estimation is

described in Appendix A.4. Two pooled machine-learning benchmarks, a support vector regression with a radial basis kernel [22] and a single-hidden-layer perceptron [23], are fitted coordinate-wise to the pooled one-step log-ratio transitions of all countries in the training window and iterated exactly like the structural map, with their hyperparameters selected by walk-forward validation inside the training window; the equal-weight combination averages the five classical benchmarks and the pooled structural model in centred log-ratio coordinates. Appendix A.4 gives the implementation details, the fallback rules and the invariance of each benchmark to the choice of log-ratio reference.

Table 1. Forecast functions, parameter vectors and free parameter counts by model and composition size. $z = \mathrm{alr}(x)$ has $m = K - 1$ coordinates; every map is iterated h times and, where it operates on z, mapped back to the simplex by the inverse transform. VAR counts refer to the conditional mean. The two machine-learning benchmarks are non-parametric maps whose hyperparameters are listed in place of a parameter vector.

Model	One-Step Map	Parameters	K = 3	K = 4
No change	$z' = z$	none	0	0
Random walk with drift	$z' = z + \delta$	δ	2	3
ARIMA, per coordinate	${z_k}'$ = ARIMA(p, d, q) in z_k	order-dependent	varies	varies
Exponential smoothing, per coordinate	${z_k}'$ = ETS with optional damped trend	order-dependent	varies	varies
VAR, lag order by AIC	$z' = c + A_1\, z + \ldots + A_p\, z_{\{-p+1\}}$	c, A_1, …, A_p	6 (p = 1)	12 (p = 1)
Dynamic Dirichlet	$x' \sim \mathrm{Dirichlet}\,(\varphi\, m(x))$, $m(x) = \mathrm{C}(\exp\,(\rho clr(x) + c))$	c, ρ, φ	4	5
Replicator–mutator, Equation (1)	$x' = F(x;\, \theta)$	α, γ, μ	4	5
Support vector regression, pooled	z' = SVR(z), radial basis kernel	$(C,\, \varepsilon) \in \{1,\, 10\} \times \{0.05,\, 0.1\}$	tuned	tuned
Perceptron, pooled	z' = MLP(z), one hidden layer	(width, penalty) $\in \{4,\, 8\} \times \{0.001,\, 0.1\}$	tuned	tuned
Equal-weight combination	clr average of five classical benchmarks and the pooled structural map	none	0	0

The parameter counts explain why the restriction is worth testing. At $K = 3$ the structural model has four parameters against six for the conditional mean of a first-order vector autoregression; at $K = 4$ it has five against twelve, because the autoregressive parameterisation is quadratic in the number of coordinates while the structural one is linear in the number of parts. On a panel of N units, the totals are $4N$ under country-specific estimation, 4 under full pooling and $4N + 4$ under the shrunk regime at $K = 3$, which is the arithmetic behind the sharing margin.

3.3. Parameter Sharing and Walk-Forward Selection

Let θ collect the parameters of Equation (1) in the unconstrained parameterisation. For a panel of units $i = 1, \ldots, N$ three regimes are considered. Under the country-specific regime, $\hat{\theta}_i$ is estimated on unit i's training window alone, so that $\mathcal{S} = \{i\}$ in Equation (3). Under the pooled regime, a single $\hat{\theta}_{\mathrm{pool}}$ is estimated on all units' training data jointly, $\mathcal{S} = \{1, \ldots, N\}$. Under the shrunk regime, the forecast for unit i uses the convex combination

$$\theta_i(\lambda) = \lambda\, \hat{\theta}_i + (1 - \lambda)\, \hat{\theta}_{\mathrm{pool}}, \qquad \lambda \in \{0, 0.5, 1\}, \tag{5}$$

with λ chosen for each unit and each origin by walk-forward validation; the training window is split into an estimation segment and a validation segment at its end, and both endpoints are estimated on the estimation segment; the candidate weights are scored

on the validation segment, and the winner is re-estimated on the full training window before forecasting. The evaluation sample is never touched. Because the combination is taken in the unconstrained parameterisation, the implied mutation rate remains in the unit interval by construction. The words are used precisely below: the pooling regime is the design margin with three settings; parameter sharing covers the two settings that estimate anything in common; full pooling is the fully pooled endpoint alone. The dynamic Dirichlet comparator is estimated under the same three regimes with the same walk-forward selection, and the drift benchmark is estimated with country-specific and with pooled drift, the latter being the grand mean of one-step log-ratio changes across all countries in the training window, so that at least one purely statistical benchmark also varies its sharing regime. The two machine-learning benchmarks exist only in the pooled regime, for the reason given in Appendix A.4.

3.4. Data and Evaluation Protocol

Three panels are constructed from Eurostat online datasets, restricted to the longest run of consecutive annual observations from 2000 onward with at least eighteen years, and with the European Union and euro area aggregates removed so that every unit is a country. The attainment panel uses population by educational attainment level for ages 25 to 64, both sexes, in per cent (Eurostat online data code edat_lfse_03), partitioned into ISCED 2011 levels 0 to 2, 3 to 4, and 5 to 8; the three parts are an exact partition. The waste panels use municipal waste by treatment operation in thousands of tonnes (Eurostat online data code env_wasmun). The three-route version partitions treated municipal waste into recycling, including composting and digestion, incineration, including energy recovery, and landfill and other disposal, which together account for 99.6% of reported treated municipal waste on average and are reclosed to sum to one within every country-year; the four-route version splits recycling into material recycling and composting with digestion, giving a four-part composition on the same countries and years and allowing the parameter-count advantage of Table 1 to be exercised. Table 2 gives the descriptives. The contrast between the panels is deliberate: attainment compositions are strongly and monotonically trending across Europe over this period, as the low-attainment share falls and the tertiary share rises, whereas municipal waste routes are noisier, subject to reporting revisions, and contain genuine structural zeros for countries that operate no incineration capacity. Zero parts are replaced multiplicatively before any log-ratio operation, with the replaced mass set to 10^{-4} in proportion units, identically for every model.

Table 2. Panel descriptives. Zero cells are parts reported as zero and subject to multiplicative replacement.

Panel	Parts	Countries	Years	Observations	Zero Cells	Origins	Target Years
Attainment	3	33	2000–2025	834	0 (0.0%)	438	2012–2025
Waste, 3 routes	3	25	2000–2024	609	111 (6.1%)	309	2012–2024
Waste, 4 routes	4	25	2000–2024	609	163 (6.7%)	309	2012–2024

The evaluation is an expanding-window rolling origin. For each unit, every model is re-estimated at every origin using only data up to that origin, starting from a training window of twelve observations, and forecasts are produced for horizons one through five, truncated at the end of the sample. This applies to the structural model in full: the shrinkage weight, the pooled parameter vector and the unit-specific parameter vector are all recomputed at every origin from training data alone, as are the hyperparameters of the machine-learning benchmarks. Two losses are computed. The primary loss is the mean absolute scaled error of Hyndman and Koehler [24], computed on log-ratio

coordinates, averaged over coordinates, and scaled by the in-sample mean absolute one-step no-change error of the corresponding training window, which makes it comparable across panels of different volatility; the secondary loss is the Aitchison distance between realised and predicted composition. Pairwise comparisons use the Diebold-Mariano statistic [25] with the small-sample correction of Harvey, Leybourne and Newbold [26] as descriptive evidence; every inferential statement rests on a bootstrap that resamples whole countries with replacement, which is the relevant clustering for a panel of national statistics, and on the model confidence set of Hansen, Lunde and Nason [27] computed at the 10% level by resampling complete country loss histories. With 25 to 33 country clusters, these *p*-values are approximations, so the substantive conclusions rest on the descriptive rankings and borderline test outcomes are treated with caution. Appendix A.5 gives the details, and Appendix A.6 tests the whole apparatus on simulated data of known provenance before it is applied to the real panels.

4. Results

4.1. One Table: The Sharing Margin Against the Model-Class Margin

Table 3 contains the whole result. It reports the mean absolute scaled error averaged over horizons one to five for every model on every panel, with the two dynamic families arranged by sharing regime, and it closes with the two margins the paper is built to compare: the reduction in error from moving a family from country-specific estimation to its best sharing regime, and the gap between the best structural specification and the best non-structural model. Figure 2 draws the two dynamic families from the same table. Horizon-by-horizon tables, the win-rate analysis, the estimated parameters and the robustness checks under the Aitchison distance, the opposite log-ratio reference and alternative zero replacements are in the Supplementary Materials; none of them changes the ordering read here.

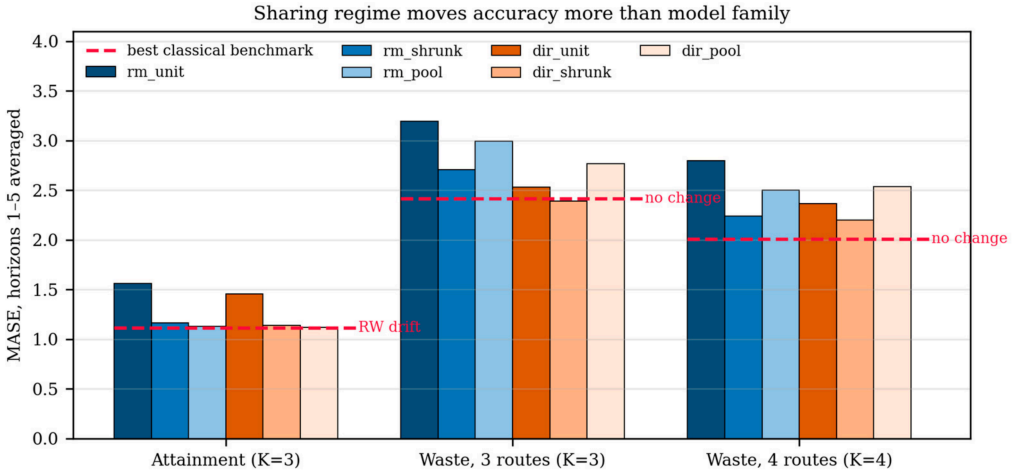


Figure 2. The two dynamic families under the three sharing regimes in each panel, mean absolute scaled error averaged over horizons one to five, with the best classical log-ratio benchmark on that panel shown as a dashed line (the random walk with drift on attainment, the no-change forecast on both waste panels). Values as in Table 3.

Three facts are visible at once. First, on every panel and in both dynamic families the country-specific regime is the worst specification of its family, and the gain from moving to the best sharing regime is large: 27.8% for the structural model on attainment, 15.3% and 19.8% on the waste panels, and 23.1%, 5.4% and 7.2% for the Dirichlet comparator. Second, the gap between the best structural specification and the best non-structural model is 1.3% on attainment and 13.1% and 11.5% on the waste panels, so on attainment the sharing

decision is worth more than twenty times the structural-versus-statistical decision, and on the waste panels the two are of comparable magnitude, with the sharing margin the larger on both. In no case is the model class the dominant consideration. Third, the best sharing regime is not the same everywhere: on attainment it is full pooling, on both waste panels it is the validation-selected shrunk specification, and full pooling there is worse than shrinkage for the structural family and worse than country-specific estimation for the Dirichlet family and for the drift benchmark. The robust finding is therefore that an appropriately selected parameter-sharing regime improves accuracy, not that full pooling is uniformly beneficial; the title's pooling is shorthand for this parameter-sharing margin.

Table 3. Mean absolute scaled error averaged over horizons one to five, all models and panels. rm and dir denote the replicator–mutator model of Equation (1) and the dynamic Dirichlet comparator; unit, pool and shrunk denote the country-specific, fully pooled and shrunk regimes of Section 3.3; rwdrift_pool is the drift benchmark with pooled drift, svr_pool and mlp_pool the pooled machine-learning benchmarks and comb the equal-weight combination. On the attainment panel, a dagger marks the nine models that the country-clustered model confidence set retains at every horizon. The sharing gain is the reduction from the country-specific regime to the better of the two sharing regimes; the model-class gap is the distance between the best structural specification and the best non-structural model on that panel.

Model	Attainment	Waste, 3 Routes	Waste, 4 Routes
rm_unit	1.564	3.196	2.796
rm_pool	1.129 †	2.992	2.503
rm_shrunk	1.167 †	2.706	2.242
dir_unit	1.455	2.530	2.368
dir_pool	1.119 †	2.769	2.537
dir_shrunk	1.139 †	2.393	2.198
Random walk with drift	1.115 †	2.780	2.363
rwdrift_pool	1.134 †	3.251	2.603
No change (naive)	2.140	2.417	2.011
ARIMA	1.151 †	2.898	2.506
Exponential smoothing	1.451	2.581	2.182
VAR	1.634	5.283	3.171
svr_pool	1.325 †	3.176	2.279
mlp_pool	1.456	3.001	2.430
comb	1.153 †	2.767	2.080
Sharing gain, replicator–mutator	27.8%	15.3%	19.8%
Sharing gain, dynamic Dirichlet	23.1%	5.4%	7.2%
Model-class gap	1.3%	13.1%	11.5%

The statistical support is as follows. On the attainment panel, the country-clustered model confidence set contains the same nine models at every horizon: the drift benchmark, the pooled drift benchmark, ARIMA, the equal-weight combination, the pooled support vector regression and the pooled and shrunk specifications of both dynamic families; in contrast, the no-change forecast, exponential smoothing, the VAR, the perceptron and both country-specific dynamic specifications are excluded throughout. Against the random walk with drift, the corrected Diebold-Mariano statistics for the shrunk structural model range from -1.45 to -0.87 across horizons, with country-clustered bootstrap p-values between 0.168 and 0.478. The structural model is therefore statistically indistinguishable from the best available forecast on this panel, without being better than it. On the waste panels, the no-change benchmark is never eliminated: it is alone in the set at horizon one on the three-route panel, the shrunk Dirichlet comparator enters from horizon two on the three-route panel and from horizon three on the four-route panel, and the sets widen with the horizon to as many as eleven of the fifteen models at horizon five, which is what resampling twenty-

five country clusters can be expected to resolve. On the three-route panel, the shrunk Dirichlet comparator finishes marginally ahead of the no-change benchmark under the primary loss, 2.393 against 2.417, but the ordering reverses under the Aitchison distance, and country-clustered tests cannot distinguish the two, so the accurate statement is parity between persistence and the best simplex-native specification rather than dominance in either direction. Against the no-change benchmark, the structural model is significantly worse at the shortest horizons on both waste panels and insignificantly worse thereafter.

4.2. Economic Reading of the Ordering

Each row of Table 3 has an economic explanation, and the explanations are the same ones that motivated the restriction in Section 2.

The attainment panel is a diffusion process of the Griliches and Fisher-Pry kind: the low-attainment share falls, and the tertiary share rises monotonically in every country, and in log-ratio coordinates the paths are close to straight lines. On such data the random walk with drift, which extrapolates the average log-ratio change, is nearly the right model, and Equation (2) explains why the pooled structural model ties it rather than beats it. The structural map is the drift benchmark plus a frequency-dependent correction $-\gamma(x_k - x_K)$ plus a shrinkage toward the barycentre; when the log-ratio paths are already straight, the correction has little to add, and the two additional parameters are paid for in estimation variance. The pooled structural model at 1.129 and the pooled drift benchmark at 1.134 sit within half a per cent of each other, as the nesting predicts. The estimated congestion parameter nevertheless carries the economics of Section 2.2: its full-sample median on the attainment panel is -0.074; it is negative for 23 of 33 countries; and for eleven of the 33 countries the estimated map admits more than one numerically located rest point, so that the same dynamics imply different long-run attainment compositions from different starting points. That is the increasing-returns story of directed technical change and peer effects, read off a forecasting model estimated for a different purpose, and the stability analysis of Appendix A.2 confirms that negative γ is exactly the condition under which the interior rest point of the map becomes unstable, and the attractors move toward the faces of the simplex. The reading is conjectural in one respect: nothing in the estimation ties γ to auxiliary country-level information such as returns to education or intergenerational mobility, and testing it against such data is a study of its own.

The waste panels are congestion processes with noise. The estimated congestion parameter is positive, with medians of 0.109 on the three-route panel and 0.244 on the four-route panel, so that treatment routes crowd as capacity constraints and rising marginal disposal costs would imply, and the interior rest point of the map is locally stable for every positive estimate, as Appendix A.2 shows. But the year-to-year movement of the route mix around that rest point is dominated by reporting revisions and by structural zeros, and on such data the no-change forecast, which is the special case of Equation (2) with no drift and no curvature, is the hardest model to beat: every dynamic model, structural or statistical, estimates a direction of movement from twelve to twenty-five noisy observations and pays for it when the direction does not materialise. The Supplementary Materials make the mechanism explicit. On the 6.1% and 8.7% of forecast events whose realised composition contains a replaced zero, persistence of the replaced value is almost exactly right and every interior-valued dynamic model pays a large log-ratio penalty; on the zero-free majority of events the ranking inverts and the shrunk Dirichlet comparator is ahead of the no-change benchmark on both panels. The no-change benchmark's lead on the waste panels is thus earned almost entirely on structural zeros, and the interesting forecasting problem on these panels is the zero-free interior, where the best simplex-native dynamic model is already ahead.

The sharing margin has the oldest explanation in forecasting: the bias-variance trade-off, with an economic gloss. At twelve to twenty-five annual observations per country, the sampling variance of four or five country-specific parameters swamps whatever heterogeneity bias pooling introduces, and the prior of Section 2.3, that the technology of adjustment is common across countries drawing on one regulatory and statistical space while the intercepts are not, is what the shrunk regime implements when it selects between the endpoints. That the Dirichlet family reproduces the sign of the effect, at a materially smaller size on the waste panels, shows that the margin belongs to the data environment rather than to the evolutionary restriction: it measures the value of information sharing in short panels, not a property of replicator dynamics. That full pooling hurts the Dirichlet model and the drift benchmark on the waste panels, and that the walk-forward selection chooses full pooling on roughly half of the unit-origins and no pooling on roughly a third, shows that the degree of sharing is a tuning decision that the data must be allowed to make, which is the subject of Section 6.

Two further results belong in the coherent set because they bound it. The machine-learning benchmarks land mid-field on every panel, ahead of country-specific structural estimation on attainment but behind the frontier of pooled specifications, and near the bottom on the three-route waste panel: the pooled sample of several hundred transitions is large enough for them to exploit but not large enough to reward flexibility over a four-parameter map. And the mutation rate is not a behavioural constant. Appendix A.3 shows that multiplicative observation noise on the simplex induces an inward drift with the same signature as mutation, so that the least-squares criterion loads measurement error onto $\hat{\mu}$, and the Supplementary Materials report that imposing $\mu = 0$ improves accuracy in every one of the nine panel-by-regime cells, by 0.9% to 17.2%, without changing any ordering. The estimated exploration rate should be read as a smoothing device selected on training data, which is the reading Section 6 builds on.

5. Discussion

5.1. The Result in the Forecasting Literature

The forecasting literature predicted the shape of this result before the data were seen, and the value of the exercise is in having measured it on compositional panels under a controlled design. Three of its strands are directly involved. The first is the long insistence, from the competitions of Makridakis and co-authors [28–30] to the survey of Petropoulos and co-authors [31], that simple methods are hard to beat and that claims of accuracy must be established with proper protocols. That an economic restriction ties the drift benchmark to trending data and loses to persistence on noisy data is neither surprising nor discreditable in that light; it is the expected outcome of testing a restriction out of sample rather than in sample.

The second strand is cross-learning and panel forecasting. The M4 and M5 competitions established that methods fitting one model to many series outperform series-by-series estimation; Montero-Manso and Hyndman [32] supply the formal argument for why global models can dominate local ones even on heterogeneous collections; Semenoglou and co-authors [33] document the accuracy of cross-learning methods; and Salinas and co-authors [34] give a prominent neural implementation. The panel-econometrics side of the same argument is older. Garcia-Ferrer, Highfield, Palm and Zellner [17] and Zellner and Hong [18] showed in the 1980s that shrinking country-specific coefficients toward a pooled mean improves international growth forecasts, and that estimator is the linear ancestor of Equation (5); Hoogstrate, Palm and Pfann [35] established that pooled estimates dominate in mean squared forecast error precisely when the time dimension is short relative to parameter heterogeneity; Baltagi [36] surveys the accumulated evidence that pooled

and shrinkage estimators outforecast their heterogeneous counterparts in short panels; Liu, Moon and Schorfheide [37] give the modern empirical-Bayes treatment; and Pesaran, Pick and Timmermann [38] show that the advantage of pooling grows with estimation uncertainty and falls with heterogeneity, and that combinations of individual and pooled forecasts are the most reliable choice across configurations. The sharing result of Table 3 is the same economics in a nonlinear, simplex-valued setting with a handful of parameters, and the simulation grid of Appendix A.6 reproduces the ordering these papers predict: under the model's own data-generating process, the best sharing regime beats country-specific estimation in every cell with homogeneous units, by 1.9% to 10.6%, while any appreciable heterogeneity hands the race to country-specific estimation at every sample length and noise level considered. The empirical dominance of sharing on the real panels is therefore not a small-sample artefact within the structural world; it indicates that the variance reduction obtained under misspecification and measurement noise outweighs a degree of parameter heterogeneity that would be decisive if the model were literally true.

The third strand is forecast combination. Timmermann [39] catalogues when combination pays and why equal weights are hard to beat; Claeskens, Magnus, Vasnev and Wang [40] show that estimating combination weights destroys the very optimality that motivates them; and Wang, Hyndman, Li and Kang [41] review the field. The Supplementary Materials report that win rates and mean losses diverge on the attainment and three-route panels, with the model that wins the most individual forecast events not being the model with the lowest mean loss, which is the classic signature of loss profiles that differ across event types and the classic argument for combination over selection. The equal-weight combination tested here finishes 3.4% behind the best single model on attainment and on the four-route waste panel and 15.6% behind on the three-route panel, and beats it nowhere, which is what the combination puzzle predicts when one of the components is already close to optimal; weighted or selective combination is a hypothesis for future work rather than a recommendation of this paper. The nonlinear and machine-learning strand completes the picture: Teräsvirta [42] concludes that carefully specified nonlinear models beat linear ones only intermittently, Zhang, Patuwo and Hu [23] catalogued how sensitive neural forecasters are to specification and sample length, Hewamalage, Bergmeir and Bandara [43] document how much data recurrent architectures need before they become competitive, and Zeng and co-authors [44] show a one-layer linear model matching transformer-based forecasters on standard benchmarks, while Medeiros, Vasconcelos, Veiga and Zilberman [45] and Smyl [46] show machine-learning methods winning when data are pooled on a scale annual official statistics cannot supply. The mid-field placement of the two pooled machine-learning benchmarks in Table 3 is consistent with every strand of that evidence.

5.2. The Result in Economics

For the evolutionary economics the restriction came from, the result has three implications. The first concerns practice. An author who estimates an evolutionary model country by country and reports in-sample fit is making the choice that the evidence identifies as most damaging to out-of-sample performance on panels like these, and the fix, a validation-selected sharing regime, improved accuracy on every panel. The compositional data analysis tradition [2,47–49] and its recent likelihood-based branch [21,50–52] supply the geometry and the coherence, but they are agnostic about where the dynamics come from; the evolutionary tradition [9–11] supplies the dynamics but has almost never asked them to forecast. The design used here joins the two under the protocol of the forecasting literature, and the joint answer is that the economics survives as interpretation while the accuracy comes from sharing.

The second implication concerns interpretation. The sign of the congestion parameter sorted the panels exactly as the theoretical examples of Section 2.2 predicted: negative under the increasing returns of directed technical change and peer effects in attainment, and positive under the congestion of capacity-constrained disposal routes; moreover, the multistability of the attainment map is the lock-in of Arthur [15] appearing in a forecasting model. These are the conclusions the applied literature draws from in-sample fits, and the point of the present design is that they survive a protocol that the fits themselves would not have survived: the estimates are interpretable even though the restriction that produced them does not forecast better than its own special cases. The mutation rate is the exception, and Appendix A.3 shows why: exploration and measurement noise leave the same second-order signature on the simplex, so no prior, penalty or profile likelihood restores a behavioural reading of $\hat{\mu}$ from public aggregates; only auxiliary knowledge of the measurement process, such as design-based sampling variances of the underlying surveys, could do so.

The third implication concerns what a negative result is for. Once the structural and statistical families are known to sit within about one per cent of each other on the attainment panel, the fifteen to twenty-eight per cent attached to the parameter-sharing decision stops being one result among many and becomes the only design choice in this environment with a first-order effect on out-of-sample accuracy, and it is the choice the applied literature currently makes by default and in the more expensive direction. A study that had found the restriction to pay would have left that margin undiscovered.

5.3. Scope and Limitations

The scope of the claim should be stated exactly, because the evidence supporting it is narrow in five respects. The panels are three, drawn from two subject domains and a single statistical producer; the frequency is annual; the country series are short, with a minimum training window of twelve observations and at most twenty-six at the final origin; the cross-section is twenty-five to thirty-three national units; and the comparator set is the classical log-ratio family, one simplex-native likelihood model and two pooled machine-learning benchmarks, with recurrent and transformer-based forecasters excluded for the scale reasons given in Appendix A.4. What the evidence supports is accordingly a conditional statement: in short annual compositional panels of official statistics, where per-unit estimation uncertainty is large relative to genuine cross-unit parameter heterogeneity, the choice of parameter-sharing regime moves out-of-sample accuracy more than the choice between an evolutionary restriction and a generic log-ratio benchmark. Appendix A.6 marks the boundary of that condition from the other side, since moderate or strong heterogeneity reverses the ranking in every cell of the grid, and Pesaran, Pick and Timmermann [38] give the analytical reason. Whether the ordering survives at higher frequency, on longer histories, in other domains or against cross-learning competitors at scale is an open empirical question, and the title of the paper names the regime studied here rather than a law of compositional forecasting.

Five limitations qualify the findings. First, and most seriously, the mutation parameter is not separately identified from observation noise, for the reason given above; for the forecasting purpose of the paper, the defect is contained, because μ enters the evaluation only as a smoothing device selected on training data and judged out of sample, and the Supplementary Materials verify that imposing $\mu = 0$ leaves every conclusion intact. Second, the waste panels contain 6.1% and 6.7% zero cells, which the multiplicative replacement handles identically for all models but which inflates every log-ratio loss and is part of why no dynamic model succeeds there; the zero-tolerant losses reported in the Supplementary Materials, computed on the raw share scale, show the pooled Dirichlet model overtaking

persistence under the Jensen-Shannon divergence on both waste panels, so the ranking of persistence against the best simplex-native model is the one conclusion that does not survive every treatment, and it is reported as parity for that reason. Third, the payoff specification is deliberately minimal; a richer structure with strategy-specific congestion or covariates entering the intercepts would add parameters and, on the evidence of Table 3, would need to be pooled to be useful, and covariates would change the exercise from unconditional to conditional forecasting. Fourth, for the minority of units with strongly negative congestion estimates, forecasts are launched inside basins of attraction whose boundaries move with the parameter estimates, so estimation error can relocate the configuration toward which the iterated map travels; the short horizons used here and the damping factor $1 - \mu$ keep the risk contained, but longer-horizon use in the increasing-returns regime should be preceded by the stability audit that Appendix A.2 makes mechanical. Fifth, the shrinkage of Equation (5) is discrete selection over a coarse grid with one weight for the whole parameter vector, not hierarchical partial pooling; a proper hierarchical estimator that shrinks each parameter separately is the first extension the next section motivates.

6. Tuning the Parameters by Walk-Forward Optimisation

The evidence of Section 4 identifies two parameters of the structural forecast function as tuning devices rather than structural constants: the shrinkage weight λ, whose selected value varies across units and origins, and the mutation rate μ, whose freely estimated value contributes variance without predictive content. A third dial, the degree of sharing applied to different blocks of the parameter vector, is implicit in the prior of Section 2.3. All three are natural objects for walk-forward optimisation, the sequential procedure in which tuning parameters are chosen on a validation segment inside the training window, applied to the next out-of-sample period, and re-chosen as the window advances. The design of this paper already uses that procedure for λ and for the hyperparameters of the machine-learning benchmarks; what follows generalises it to the structural model and states what the evidence says about each dial.

The procedure is as follows. At origin t_0 the training window of unit i is split into an estimation segment and a validation segment consisting of its last q years; the rule used for the machine-learning benchmarks here, a quarter of the window with a minimum of two years, is a reasonable default. On the estimation segment, the country-specific and pooled parameter vectors are estimated by Equation (3) for every candidate value of the tuning vector (λ, μ), with λ on a grid in the unit interval and μ either fixed at zero, fixed at a grid value, or freely estimated. Each candidate is scored on the validation segment by the multi-horizon loss actually used for evaluation, the scaled error averaged over the horizons the forecaster will be asked to produce, so that the selection criterion and the evaluation criterion coincide. The winning candidate is then re-estimated on the full training window and used to forecast horizons one to five from t_0; the window is advanced by one year, and the procedure repeats, so that the tuning vector is re-optimised at every origin and no forecast uses information beyond its origin. Because the search is nested inside the rolling origin, the resulting accuracy is an honest out-of-sample measure of the tuned forecaster, and because the same nesting is applied to every competitor, the comparison remains fair.

The evidence bounds what each dial can deliver. For the shrinkage weight, the walk-forward selection over the coarse grid $\{0, 0.5, 1\}$ chose full pooling on 49.1%, 49.5% and 50.5% of the unit-origins on the three panels, no pooling on 27.4%, 32.4% and 31.7%, and the intermediate weight on the remaining 23.5%, 18.1% and 17.8%. Refining the grid to eleven points in the unit interval moved accuracy by a few percent and in both directions, from 2.706 to 2.621 and from 2.242 to 2.216 on the waste panels but from 1.167 to 1.206 on attainment, and it redistributed weight away from the corners without eliminating them,

the endpoints being chosen in 58% to 66% of unit-origins against 77% to 82% under the coarse grid; the selected weight also switched between adjacent origins more often under the fine grid, 36% to 44% of the time against 17% to 23%. The lesson for a walk-forward implementation is that the candidate set should be kept coarse when the validation segment is short, because a rich menu adds selection noise that a validation window of a few years cannot resolve, and that the corner frequencies describe what such a window can resolve rather than a natural division of countries into homogeneous and heterogeneous types.

For the mutation rate, the evidence is one-sided. Fixing $\mu = 0$ improved accuracy in all nine panel-by-regime cells, by 0.9% to 17.2%, with the largest gains for the country-specific regime on the waste panels, exactly where the recovery experiment of Appendix A.6 predicts the estimated rate to be least reliable; imposing the cross-country median instead helped the country-specific regime but hurt the pooled one, whose own estimated smoothing level is smaller. A walk-forward search should therefore always include $\mu = 0$ in the candidate set and treat the freely estimated rate as one candidate among others, and it should expect the selection to favour zero or a small fixed value on panels this short. The third dial, parameter-specific sharing, is not implemented here and is the natural next step: the prior of Section 2.3 suggests pooling the adjustment parameters γ and μ while allowing the intercepts to vary, which a walk-forward search can test by letting the shrinkage weight differ between the two blocks, and a hierarchical random-coefficient formulation of the map with second-level variances estimated across countries would replace the grid search with a proper estimator. Two cautions apply to any such extension. The validation segment must be taken from the end of the training window, never from its interior, because the losses that matter are those of forecasts launched from the most recent state; and an expanding window is preferable to a fixed-length rolling window on panels of twelve to twenty-six observations, because the variance cost of discarding early observations dominates the bias cost of structural change at these lengths, though a fixed-length window becomes the right choice once the history is long enough for the trade-off to reverse.

7. Conclusions

We specified a discrete-time replicator–mutator map as a forecast function for compositional time series, showed that it nests the random walk with drift and the no-change forecast as special cases, estimated it under three parameter-sharing regimes, and evaluated it, together with a non-evolutionary dynamic Dirichlet comparator under the same regimes, against classical log-ratio benchmarks, a pooled drift benchmark, two pooled machine-learning benchmarks and an equal-weight combination on three Eurostat panels with a rolling-origin design, scale-free losses, country-clustered tests of equal predictive ability and model confidence sets.

The evolutionary restriction does not pay for itself in accuracy. It matches the best benchmark on educational attainment, where it enters the model confidence set at every horizon, and it loses to a no-change forecast on municipal waste routes, as does every other dynamic model except the shrunk dynamic Dirichlet comparator, which reaches parity there. What does pay is parameter sharing: moving from country-specific estimation to the best sharing regime, full pooling on attainment and validation-selected shrinkage on the waste panels, reduces scaled error by 15.3% to 27.8% across the three panels, an effect larger than the one separating the best structural model from the best benchmark, and it reappears with the same sign in a non-evolutionary family. The estimated congestion parameters nevertheless sort the panels as the economics predicts, with increasing returns in attainment and congestion in waste routes, so the restriction earns its keep as interpretation rather than as accuracy.

https://doi.org/10.3390/forecast8050089

For applied researchers using evolutionary dynamics on panels of official statistics, the operative message is that the parameter-sharing decision deserves at least as much attention as the specification of payoffs, that the shrinkage weight and the mutation rate should be selected by walk-forward optimisation inside the training window rather than fitted in sample, and that out-of-sample evaluation should be the standard rather than the exception. Two boundaries travel with that message. The comparison is against classical log-ratio benchmarks, one simplex-native likelihood model and two pooled machine-learning benchmarks rather than against cross-learning methods at the scale of long and wide corpora, and the panels are short, annual and drawn from a single statistical producer; a simulation grid shows the ordering reversing once cross-country parameter heterogeneity is appreciable. Within those boundaries the ordering is clear, tested and reproducible; outside them it is a hypothesis. The contribution is accordingly not the failure of one restriction but the relocation of the design margin: in this data environment the question worth asking of a compositional forecasting model is not which dynamics it imposes, but how much of its parameter vector is estimated in common across units.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/forecast8050089/s1, Figure S1: Mean absolute scaled error by forecast horizon, educational attainment panel, 33 countries, expanding-window rolling origin; Figure S2: Mean absolute scaled error by forecast horizon, three-route municipal waste panel, 25 countries; Figure S3: Share of forecast events on which each model attains the lowest scaled error, by panel, fifteen-model set; Figure S4: Full-sample estimates of the congestion parameter and the exploration parameter, one point per country, educational attainment panel; Table S1: Mean absolute scaled error by horizon, educational attainment panel; Table S2: Mean absolute scaled error averaged over horizons one to five, municipal waste panels, complete model set; Table S3: Win rates: share of forecast events on which each model attains the lowest scaled error, fifteen-model set; Table S4: Stability audit of rest points located at the full-sample estimates; Table S5: Mean absolute scaled error averaged over horizons one to five under three treatments of the mutation parameter; Table S6: Mean Aitchison distance, all horizons pooled, baseline zero replacement; Table S7: Mean absolute scaled error, all horizons, with the first part as the additive log-ratio reference; Table S8: Zero-replacement sensitivity: mean absolute scaled error, all horizons, waste panels (W3, W4) under replacement values 10^{-3}, 10^{-4} (baseline) and 10^{-5}.

Funding: This research received no external funding.

Data Availability Statement: The study uses publicly available Eurostat data (online data codes edat_lfse_03 and env_wasmun). The estimation and evaluation code, the constructed panels, and all result tables accompany the submission as a replication package.

Conflicts of Interest: The author declares no conflicts of interest.

Appendix A. Technical Details

Appendix A.1. Simplex Geometry and Log-Ratio Coordinates

A composition with K parts is a vector $x = (x_1, \ldots, x_K)$ with $x_k > 0$ whose parts sum to one; the set of such vectors is the open unit simplex. Because the parts are tied together by the sum constraint, the simplex is not a vector space under ordinary addition, and correlations between raw parts are distorted by construction. The standard solution, due to Aitchison [1,2], is to move to log-ratio coordinates. The additive log-ratio transform takes one part as the reference,

$$\mathrm{alr}(x) = \left(\ln \frac{x_1}{x_K}, \ldots, \ln \frac{x_{K-1}}{x_K} \right), \tag{A1}$$

while the centred log-ratio transform divides by the geometric mean $g(x) = \left(\prod_{j=1}^{K} x_j\right)^{1/K}$,

$$\mathrm{clr}(x) = \left(\ln\frac{x_1}{g(x)}, \ldots, \ln\frac{x_K}{g(x)}\right). \tag{A2}$$

Both transforms are invertible up to closure, the renormalisation $C(z) = z/\sum_j z_j$ that returns a positive vector to the simplex; an isometric variant exists [53], but is not needed here. The Aitchison distance between two compositions is the Euclidean distance between their clr images, and it is the metric in which the estimator of Equation (3) and the secondary loss are defined. Throughout, the statistical benchmarks operate on alr coordinates with the last part of each composition as reference, ISCED 5–8 for attainment and landfill for both waste panels, and are mapped back through the inverse transform. The modern treatment of the geometry is Pawlowsky-Glahn, Egozcue and Tolosana-Delgado [47]; applications to time series include the vector autoregressions of Kynčlová, Filzmoser and Hron [48] and the state space approach of Snyder and co-authors [49], which belongs to the same innovations family as the exponential smoothing benchmark. Hierarchical reconciliation [54–56] pursues coherence with respect to aggregation constraints rather than the sum-to-one constraint; the two problems are cousins, but reconciliation operates on unbounded totals while compositional forecasting operates inside the simplex, which is why the benchmark set is built from log-ratios rather than reconciliation methods.

Appendix A.2. Rest Points, Mutation and Local Stability

For $\mu = 0$ the vertices of the simplex are rest points of the map of Equation (1), and the long-horizon forecast of a pure replicator model is generically a corner. For $\mu > 0$ this is no longer possible: applying the map at a vertex e_k returns $(1-\mu)e_k + \mu/K \neq e_k$, so no vertex is a rest point, and as $\mu \to 1$ the map collapses to the constant map at the barycentre. The mutation parameter therefore interpolates between the selection attractor and the centre of the simplex, and in forecasting terms it operates as a shrinkage device: each application of the map injects weight μ toward the uniform composition, so over a five-step forecast the cumulative weight of the injections is $1-(1-\mu)^5$, which is 8.7% at the median attainment estimate of $\mu = 0.018$ and about 3% at the waste medians. Numerically, with $K = 3$, $\alpha = (0.4, -0.2, 0)$ and $\gamma = 1.5$, the interior rest point moves from $(0.5556, 0.1556, 0.2889)$ at $\mu \approx 0$ to $(0.5467, 0.1658, 0.2875)$ at $\mu = 0.02$, a small displacement toward the barycentre at one step and a large one at long horizons, since it is the limit point rather than the local slope that mutation relocates.

Write the pure selection step as $y(x)$ with components $y_k = x_k e^{\pi_k(x)}/\sum_j x_j e^{\pi_j(x)}$ and $\pi_k(x) = \alpha_k - \gamma x_k$, so that the full map is $F(x) = (1-\mu)y(x) + \mu/K$. Differentiating $\ln y_k$ and using $\partial\pi_k/\partial x_l = -\gamma\delta_{kl}$ gives

$$\frac{\partial y_k}{\partial x_l} = y_k\left[\delta_{kl}\left(\frac{1}{x_l} - \gamma\right) - \frac{y_l}{x_l}(1 - \gamma x_l)\right]. \tag{A3}$$

Because the mutation part of F is affine with a constant term μ/K, the Jacobian of the full map is exactly $DF(x) = (1-\mu)Dy(x)$ at every point of the simplex, so mutation damps the spectrum of every rest point by the factor $1-\mu$. Local stability of a rest point x^* is governed by the spectral radius of $DF(x^*)$ restricted to the tangent space of the simplex: the point attracts nearby trajectories if the radius is below one and repels them along some direction if it exceeds one. At an interior rest point of the pure selection step, payoffs are equalised and $y(x^*) = x^*$, so the expression above collapses to

$$Dy(x^*) = I - \gamma M(x^*), \qquad M(x) = \mathrm{diag}(x) - xx^{\top}. \tag{A4}$$

The matrix $M(x)$ is the covariance matrix of a categorical random variable with probabilities x; on the tangent space it is positive definite with eigenvalues $\lambda_i \in (0,1)$. The eigenvalues of $DF(x^*)$ are therefore $(1-\mu)(1-\gamma\lambda_i)$, up to terms of order μ arising from the displacement of the rest point itself. For $\gamma > 0$, the interior rest point is locally stable whenever $\gamma\lambda_{\max} < 2$, and since $\lambda_{\max} < 1$, any congestion parameter below two guarantees stability, far above the positive estimates reported in the Supplementary Materials. For $\gamma < 0$, every eigenvalue equals $(1-\mu)(1+|\gamma|\lambda_i)$ and the largest exceeds one for all but implausibly large mutation rates, so the interior point repels, the attractors lie toward the faces of the simplex, and they need not be unique; this is the analytical counterpart of the multistability reported in Section 4.2. At the parameter values of the worked example above, the tangent-space eigenvalues at the interior rest point are 0.714 and 0.410 for $\mu = 0$, both inside the unit circle, and numerical differentiation at $\mu = 0.02$ returns 0.675 and 0.405, confirming both the closed form and the damping. For any estimated unit, stability of a located rest point can be audited mechanically by evaluating the Jacobian formula at the numerically located point; the audit of every rest point located at the full-sample estimates is reported in the Supplementary Materials, where all located points are locally stable, as they must be because the location algorithm finds attractors by iterating the map forward, and where the attainment attractors are seen to contract slowly, with a median tangent-space spectral radius of 0.962, consistent with the slow monotone drift of the attainment data. At every located point, the closed form agrees with numerical differentiation of the map to at least eight decimal places.

Appendix A.3. Estimation and the Identification of the Mutation Parameter

The objective of Equation (3) is minimised by a bounded quasi-Newton routine from twelve starting values, of which one is the no-selection, no-congestion point and the remainder are drawn at random, and the best local optimum is retained. The logistic reparameterisation of μ, combined with box bounds on all parameters, acts as an implicit regulariser: it keeps the mutation rate strictly inside the unit interval at every origin and excludes the degenerate corner solutions that unconstrained estimation would admit. Zero parts are replaced multiplicatively before any log-ratio operation, with the replaced mass set to 10^{-4} in proportion units and the remaining parts rescaled so that the composition still closes to one; the same replacement is applied identically to every model, so it cannot favour one competitor over another, and the Supplementary Materials examine the sensitivity of every waste-panel conclusion to this value and report two zero-tolerant losses computed on the raw share scale.

The mutation parameter is not separately identified from observation noise, and the reason can be made precise. Let the observed composition be $\widetilde{x} = C(x \odot e^{\varepsilon})$, where C is the closure operation and ε is mean-zero noise with standard deviation σ on log coordinates. A second-order expansion gives

$$\mathbb{E}\left[\widetilde{x}_k\right] - x_k = \sigma^2 x_k \left(\sum_{j=1}^{K} x_j^2 - x_k \right) + o\left(\sigma^2\right), \tag{A5}$$

an inward drift that vanishes at the barycentre and moves mass from over-represented toward under-represented parts. The mutation flow $\mu(1/K - x_k)$ has exactly the same two properties, so at the sample sizes that official annual statistics deliver, the least squares criterion cannot separate the two drifts and loads the noise-induced one onto $\hat{\mu}$. This also explains why penalised or Bayesian estimation of μ, while numerically stabilising, cannot restore a structural reading: a prior on μ is a choice of how much inward smoothing to impose, not information that distinguishes exploration from noise. The one route to

structural identification is auxiliary knowledge of the measurement process itself, such as design-based sampling variances of the underlying surveys. The recovery experiment of Appendix A.6 documents the consequence: the estimated μ is biased upward in every configuration, and it should be read as a reduced-form smoothing parameter selected on training data, which is how Section 6 treats it.

Appendix A.4. Benchmark Implementation

Univariate ARIMA models are fitted per log-ratio coordinate with the order selected from a curated grid by corrected Akaike information; exponential smoothing is fitted per coordinate with and without a damped additive trend, again selected by corrected Akaike information [19]; the vector autoregression is fitted to the full log-ratio vector with lag order chosen by Akaike information. Whenever a training window is too short or an optimiser fails, the routine falls back to the random walk with drift. At these sample sizes, the ARIMA and exponential smoothing fits never fall back, while the VAR falls back to the drift model on the four-route panel in 32% of unit-origins, exactly the training windows shorter than the sixteen observations its short-sample rule requires, and never on the three-part panels, so the reported four-route VAR is a mixture of VAR and drift forecasts on the shortest windows. The no-change and drift benchmarks and the VAR are invariant to the choice of reference part, the first two because their forecasts are built from the last observation and the mean log-ratio change, which map exactly across bases, and the VAR because unrestricted least squares is equivariant under nonsingular linear transformations of the coordinate vector; only the univariate ARIMA and exponential smoothing fits depend on the reference part, and the Supplementary Materials re-estimate both with the opposite reference and show that every conclusion survives.

The dynamic Dirichlet comparator specifies the observed composition as Dirichlet distributed about a mean that follows a first-order recursion linear in the Aitchison geometry,

$$m(x) = C(\exp(\rho \operatorname{clr}(x) + c)), \qquad x_{t+1} \mid x_t \sim \text{Dirichlet}(\varphi\, m(x_t)), \tag{A6}$$

with $K - 1$ free coordinates in c, a scalar persistence parameter ρ and a precision φ, hence $K + 1$ free parameters. The recursion contains no frequency-dependent selection, and because ρ is a scalar, the model is invariant to the choice of log-ratio basis. Parameters are estimated by maximum likelihood, point forecasts iterate the mean recursion, and the model is evaluated under the same three sharing regimes, rolling origins, losses and zero treatment as the structural model, with the shrinkage weight selected by the same walk-forward hold-out. Like the structural model with positive mutation, the Dirichlet transition places no mass on exact zeros, a support limitation common to simplex-native specifications.

The two machine-learning benchmarks are included as the members of that family the pooled sample can support. The first, svr_pool, is a support vector regression with a radial basis kernel [22], fitted coordinate-wise to the pooled one-step log-ratio transitions of all countries in the training window with standardised inputs and outputs; the second, mlp_pool, is a single-hidden-layer perceptron with hyperbolic tangent activation [23] fitted to the same pooled transitions. Each is a one-step map iterated for multi-step forecasts, exactly parallel to the structural and Dirichlet maps. Their hyperparameters, the regularisation constant and tube width of the support vector regression over $\{1, 10\} \times \{0.05, 0.1\}$ and the hidden width and penalty of the perceptron over $\{4, 8\} \times \{0.001, 0.1\}$, are selected on a temporal hold-out inside the training window, the last quarter of the successor years with a minimum of two, and the winning configuration is refit on the full pooled window, so that no forecast uses information beyond its origin. Country-specific versions are deliberately omitted: with at most twenty-five transitions per country they would be under-identified, whereas pooled across thirty-three countries the training sample

at a late origin is of the order of several hundred transitions in two log-ratio coordinates, against which a support vector regression on a few lags or a perceptron with a small number of units is estimable. Both benchmarks were estimated successfully at every unit-origin on all three panels; they share the basis dependence of the univariate benchmarks and were not re-estimated under the opposite reference, since they sit far enough from the frontier on every panel that a basis effect of the size found for ARIMA and exponential smoothing would not change their placement. Recurrent and transformer-based forecasters are outside the comparator set for reasons of scale: a single long short-term memory cell with eight hidden units already carries more than three hundred weights, which exceeds the transitions available to any one country by more than an order of magnitude, and the architectures reported in that literature are trained on series two to three orders of magnitude longer than these [43,44]. No conclusion depends on where that boundary sits, because the parameter-sharing margin is measured within families, so adding or removing a family moves only the model-class margin, which is already small.

The pooled drift benchmark uses as drift the grand mean of one-step log-ratio changes across all countries in the training window. The equal-weight combination averages the five classical benchmarks and the pooled structural model in centred log-ratio coordinates and closes the result to the simplex.

Appendix A.5. Losses, Tests and the Model Confidence Set

The scaled error is computed on additive log-ratio coordinates and, unlike the Aitchison distance, is not invariant to the log-ratio basis; this is one reason both losses are reported, and the Supplementary Materials repeat the full accuracy comparison under the Aitchison distance, under the opposite additive log-ratio reference, and under two zero-tolerant losses computed on the raw share scale, the mean absolute error across shares and the Jensen-Shannon divergence. Pairwise comparisons use the Diebold-Mariano statistic [25] with the small-sample correction of Harvey, Leybourne and Newbold [26]; the conditional version of Giacomini and White [57] and the survey of Elliott and Timmermann [58] frame the protocol. The statistic for a given horizon is computed on the loss-differential series pooled across all unit-origin forecast events at that horizon, stacked country by country in origin order. Within a country, consecutive origins share overlapping multi-step forecast windows, and the rectangular kernel with lags up to one less than the horizon absorbs that overlap-induced serial correlation; the ordering across country boundaries, by contrast, is arbitrary, so the kernel is a within-country overlap correction rather than a full treatment of the panel dependence structure. The Harvey-Leybourne-Newbold factor rescales the statistic, and p-values use a t distribution with the event count minus one degrees of freedom. For this reason, the corrected Diebold-Mariano statistics are reported as descriptive evidence, and every inferential statement rests on the country-clustered bootstrap, which collapses loss differentials to country means before resampling countries with replacement, with 2000 centred draws, and which therefore preserves within-country serial dependence and cross-origin overlap by construction. The model confidence set of Hansen, Lunde and Nason [27] is computed in its range-statistic form at the 10% level by resampling complete country loss histories, with 1000 draws, so that the same clustering underlies every statement about set membership; an event-level version that resamples individual unit-origin observations was retained only as a descriptive sensitivity check, and no conclusion of the paper rests on it. Figure A1 assembles the whole procedure in a single diagram.

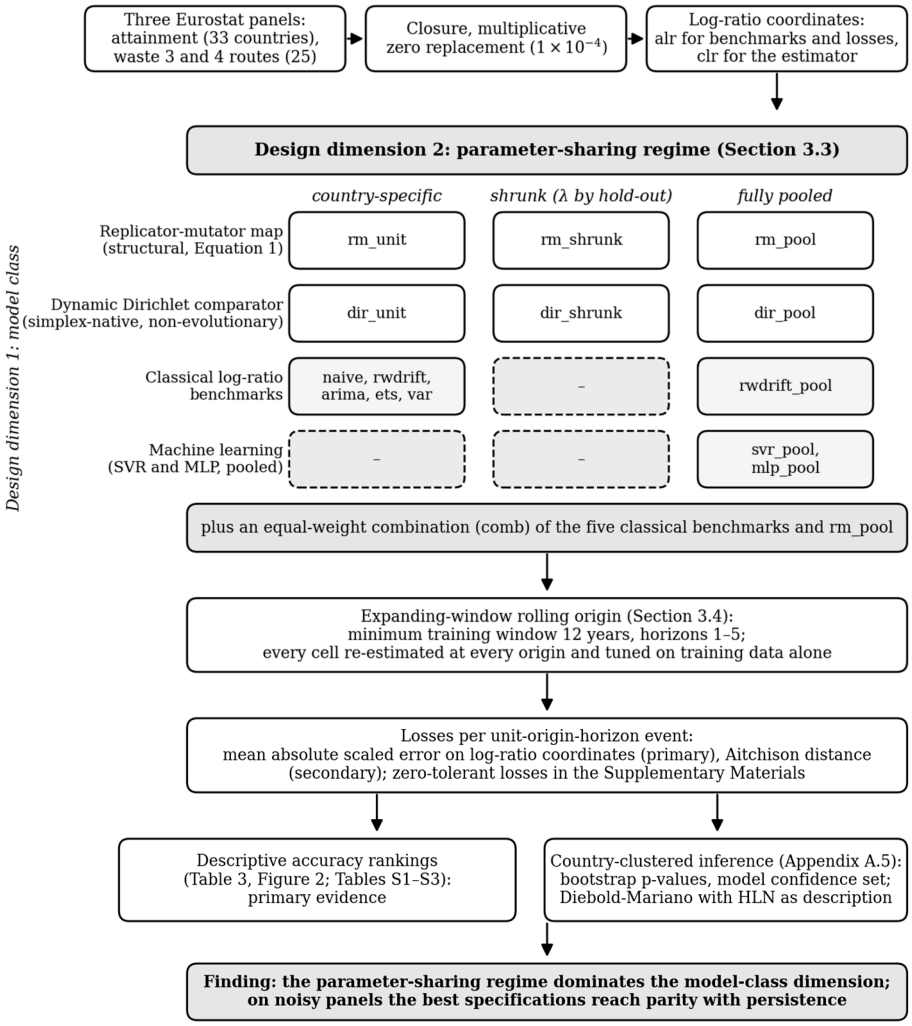


Figure A1. The research design. The study crosses a model-class dimension, running from the structural replicator–mutator map through the dynamic Dirichlet comparator and the classical log-ratio benchmarks to two pooled machine-learning benchmarks, with a parameter-sharing dimension with country-specific, shrunk and fully pooled regimes. Comparisons within a column identify the model-class margin and comparisons within a row the parameter-sharing margin. Every cell passes through the same expanding-window rolling origin, is estimated and tuned on training data alone, and is scored and compared by the protocol of Appendix A, with descriptive rankings as the primary evidence and country-clustered inference as support.

Appendix A.6. Monte Carlo Evidence

Panels of forty units were generated from the structural map with logistic-normal observation noise on log-ratio coordinates, varying the sample length and the noise standard deviation. Table A1 reports the bias and root mean squared error of the estimated congestion and mutation parameters. The congestion parameter is recovered well: at low noise the median relative error is within a third of one per cent of zero at every sample length, and the large root mean squared errors at higher noise are driven by a small number of units for which the objective is nearly flat, not by a systematic tilt. The mutation parameter is a different matter. Its bias is positive in every configuration, and its median relative error deteriorates sharply with noise, reaching a factor of seven at the noisiest configuration, which is the confounding of Appendix A.3 at work. Table A1 also doubles as guidance on minimum sample sizes. Country-specific estimation of the four-parameter model is reliable at $T = 15$ only under implausibly clean data; at realistic noise levels the congestion

parameter needs forty or more observations, and the mutation parameter is compromised at every length considered. Official annual statistics deliver training windows of twelve to twenty-five observations, which places applied work squarely in the regime where the simulations predict unit-level estimation to be dominated by variance, and this is what Section 4 finds on the real panels.

Table A1. Recovery of the congestion parameter γ and the mutation parameter μ. Forty units per configuration, three parts, data generated from the structural map with logistic-normal observation noise of standard deviation σ on log-ratio coordinates.

T	σ	Bias γ	RMSE γ	Med. Rel. Err. γ	Bias μ	RMSE μ	Med. Rel. Err. μ
15	0.02	0.008	0.116	−0.003	0.007	0.023	0.055
15	0.05	−0.770	4.045	−0.015	0.098	0.231	0.470
15	0.10	−1.061	3.652	−0.086	0.174	0.311	1.363
25	0.02	−0.107	0.309	−0.013	0.085	0.185	0.623
25	0.05	−0.033	0.477	0.006	0.083	0.188	0.062
25	0.10	−0.881	2.217	−0.193	0.253	0.360	7.120
40	0.02	−0.062	0.198	−0.003	0.064	0.144	0.113
40	0.05	−0.180	1.090	0.002	0.130	0.247	0.304
40	0.10	0.174	0.577	0.125	0.099	0.179	2.014

Two further experiments run the full evaluation machinery on generated data. In the first, twelve units of twenty-six annual observations are generated from the structural map with parameters that differ across units, so that the panel is heterogeneous and noisy rather than a best case in which sharing is trivially favoured. In the second, they are generated from a log-ratio VAR(1) with drift, a process outside the structural family. Table A2 reports the results. Under correct specification, the country-specific structural model wins, as expected, and the fully pooled version is badly hurt because the generated units have heterogeneous parameters. Under misspecification, the VAR wins, again as expected, and the structural model degrades but does not collapse: the shrunk version remains ahead of ARIMA, of the random walk with drift, and of both other structural regimes. These experiments establish that the machinery is wired correctly and that the sharing regime is capable of moving accuracy in either direction depending on the true degree of heterogeneity.

Table A2. Monte Carlo horse races, mean absolute scaled error averaged over horizons one to five. Panel A generates data from the structural map; Panel B from a log-ratio VAR(1) with drift. Model abbreviations as in Table 3.

Model	Panel A: Correct	Panel B: Misspecified
rm_unit	0.489	1.355
rm_shrunk	0.533	1.232
rm_pool	0.833	1.432
Naive	0.586	1.086
Random walk with drift	1.768	1.440
ARIMA	1.287	1.332
Exponential smoothing	0.689	1.083
VAR	0.514	0.977

A third experiment maps the boundary between those directions. Panels of twelve units are generated from the structural map on a grid crossing the sample length (T of 15, 25 and 40), the degree of between-unit parameter heterogeneity (none, moderate and strong) and the observation-noise level (σ of 0.02 and 0.10), and the full rolling-origin evaluation,

including the Dirichlet comparator and the pooled drift benchmark, is run in every cell. Table A3 reports the grid. The sign of the sharing gain is governed by heterogeneity rather than by sample length or noise alone: with homogeneous units, the gain of the best sharing regime over country-specific estimation is positive in every cell, from 1.9% to 10.6%, while moderate or strong heterogeneity makes country-specific estimation best in every cell, with the discrete shrinkage selection of Equation (5) tracking the better endpoint closely. The Dirichlet comparator and the pooled drift benchmark are dominated whenever the structural model is true, as expected under correct specification, so the family margin and the sharing margin are separately visible. The grid thereby bounds the empirical claim of Section 4: sharing dominates where effective heterogeneity is small relative to noise and misspecification, which is where the Eurostat panels evidently sit.

Table A3. Monte Carlo grid: mean absolute scaled error by cell (twelve units per cell, structural data-generating process, observation-noise level σ). Gain is the improvement of the better sharing regime over country-specific estimation.

T	Het.	σ	rm_unit	rm_pool	rm_shrunk	dir_pool	rwdrift_pool	Gain (%)
15	none	0.02	0.123	0.110	0.111	1.279	1.543	10.6
15	none	0.10	0.361	0.327	0.337	1.015	1.513	9.4
15	moderate	0.02	0.326	1.482	0.356	1.350	1.424	−9.2
15	moderate	0.10	0.717	1.138	0.870	1.191	1.021	−21.3
15	strong	0.02	0.162	1.881	0.168	1.927	1.601	−3.7
15	strong	0.10	0.521	0.818	0.548	0.945	0.808	−5.2
25	none	0.02	0.359	0.352	0.360	0.360	1.035	1.9
25	none	0.10	0.581	0.544	0.572	0.824	1.642	6.4
25	moderate	0.02	0.253	3.003	0.253	3.246	2.940	0.0
25	moderate	0.10	0.690	1.455	0.739	1.675	1.189	−7.1
25	strong	0.02	0.315	0.923	0.325	0.585	0.875	−3.2
25	strong	0.10	0.518	1.687	0.562	1.947	1.664	−8.5
40	none	0.02	0.352	0.339	0.356	0.394	2.085	3.7
40	none	0.10	0.632	0.604	0.617	0.753	1.322	4.4
40	moderate	0.02	0.200	2.299	0.200	2.619	2.435	0.0
40	moderate	0.10	0.629	1.194	0.674	1.188	1.123	−7.2
40	strong	0.02	0.233	2.000	0.233	2.624	1.866	0.0
40	strong	0.10	0.544	1.563	0.570	1.552	1.460	−4.8

References

1. Aitchison, J. The statistical analysis of compositional data. *J. R. Stat. Soc. Ser. B (Methodol.)* **1982**, *44*, 139–177. [CrossRef]
2. Aitchison, J. *The Statistical Analysis of Compositional Data*; Chapman and Hall: London, UK, 1986.
3. Taylor, P.D.; Jonker, L.B. Evolutionarily stable strategies and game dynamics. *Math. Biosci.* **1978**, *40*, 145–156. [CrossRef]
4. Hofbauer, J.; Sigmund, K. *Evolutionary Games and Population Dynamics*; Cambridge University Press: Cambridge, UK, 1998.
5. Weibull, J.W. *Evolutionary Game Theory*; MIT Press: Cambridge, MA, USA, 1995.
6. Sandholm, W.H. *Population Games and Evolutionary Dynamics*; MIT Press: Cambridge, MA, USA, 2010.
7. Schlag, K.H. Why imitate, and if so, how? A boundedly rational approach to multi-armed bandits. *J. Econ. Theory* **1998**, *78*, 130–156.
8. Nowak, M.A.; Komarova, N.L.; Niyogi, P. Evolution of universal grammar. *Science* **2001**, *291*, 114–118. [CrossRef] [PubMed]
9. Komarova, N.L. Replicator-mutator equation, universality property and population dynamics of learning. *J. Theor. Biol.* **2004**, *230*, 227–239. [CrossRef] [PubMed]
10. Fudenberg, D.; Levine, D.K. *The Theory of Learning in Games*; MIT Press: Cambridge, MA, USA, 1998.
11. Young, H.P. *Individual Strategy and Social Structure: An Evolutionary Theory of Institutions*; Princeton University Press: Princeton, NJ, USA, 1998.
12. Nelson, R.R.; Winter, S.G. *An Evolutionary Theory of Economic Change*; Belknap Press of Harvard University Press: Cambridge, MA, USA, 1982.
13. Griliches, Z. Hybrid corn: An exploration in the economics of technological change. *Econometrica* **1957**, *25*, 501–522. [CrossRef]

14. Fisher, J.C.; Pry, R.H. A simple substitution model of technological change. *Technol. Forecast. Soc. Change* **1971**, *3*, 75–88. [CrossRef]
15. Arthur, W.B. Competing technologies, increasing returns, and lock-in by historical events. *Econ. J.* **1989**, *99*, 116–131. [CrossRef]
16. Acemoglu, D. Why do new technologies complement skills? Directed technical change and wage inequality. *Q. J. Econ.* **1998**, *113*, 1055–1089. [CrossRef]
17. Garcia-Ferrer, A.; Highfield, R.A.; Palm, F.; Zellner, A. Macroeconomic forecasting using pooled international data. *J. Bus. Econ. Stat.* **1987**, *5*, 53–67. [CrossRef]
18. Zellner, A.; Hong, C. Forecasting international growth rates using Bayesian shrinkage and other procedures. *J. Econom.* **1989**, *40*, 183–202. [CrossRef]
19. Hyndman, R.J.; Koehler, A.B.; Ord, J.K.; Snyder, R.D. *Forecasting with Exponential Smoothing: The State Space Approach*; Springer: Berlin, Germany, 2008.
20. Grunwald, G.K.; Raftery, A.E.; Guttorp, P. Time series of continuous proportions. *J. R. Stat. Soc. Ser. B* **1993**, *55*, 103–116. [CrossRef]
21. Katz, H.; Brusch, K.T.; Weiss, R.E. A Bayesian Dirichlet auto-regressive moving average model for forecasting lead times. *Int. J. Forecast.* **2024**, *40*, 1556–1567. [CrossRef]
22. Sapankevych, N.I.; Sankar, R. Time series prediction using support vector machines: A survey. *IEEE Comput. Intell. Mag.* **2009**, *4*, 24–38. [CrossRef]
23. Zhang, G.; Patuwo, B.E.; Hu, M.Y. Forecasting with artificial neural networks: The state of the art. *Int. J. Forecast.* **1998**, *14*, 35–62.
24. Hyndman, R.J.; Koehler, A.B. Another look at measures of forecast accuracy. *Int. J. Forecast.* **2006**, *22*, 679–688. [CrossRef]
25. Diebold, F.X.; Mariano, R.S. Comparing predictive accuracy. *J. Bus. Econ. Stat.* **1995**, *13*, 253–263. [CrossRef]
26. Harvey, D.; Leybourne, S.; Newbold, P. Testing the equality of prediction mean squared errors. *Int. J. Forecast.* **1997**, *13*, 281–291. [CrossRef]
27. Hansen, P.R.; Lunde, A.; Nason, J.M. The model confidence set. *Econometrica* **2011**, *79*, 453–497. [CrossRef]
28. Makridakis, S.; Spiliotis, E.; Assimakopoulos, V. Statistical and Machine Learning forecasting methods: Concerns and ways forward. *PLoS ONE* **2018**, *13*, e0194889. [CrossRef] [PubMed]
29. 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]
30. Makridakis, S.; Spiliotis, E.; Assimakopoulos, V. M5 accuracy competition: Results, findings, and conclusions. *Int. J. Forecast.* **2022**, *38*, 1346–1364. [CrossRef]
31. 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]
32. Montero-Manso, P.; Hyndman, R.J. Principles and algorithms for forecasting groups of time series: Locality and globality. *Int. J. Forecast.* **2021**, *37*, 1632–1653. [CrossRef]
33. Semenoglou, A.-A.; Spiliotis, E.; Makridakis, S.; Assimakopoulos, V. Investigating the accuracy of cross-learning time series forecasting methods. *Int. J. Forecast.* **2021**, *37*, 1072–1084. [CrossRef]
34. Salinas, D.; Flunkert, V.; Gasthaus, J.; Januschowski, T. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *Int. J. Forecast.* **2020**, *36*, 1181–1191. [CrossRef]
35. Hoogstrate, A.J.; Palm, F.C.; Pfann, G.A. Pooling in dynamic panel-data models: An application to forecasting GDP growth rates. *J. Bus. Econ. Stat.* **2000**, *18*, 274–283. [CrossRef]
36. Baltagi, B.H. Forecasting with panel data. *J. Forecast.* **2008**, *27*, 153–173. [CrossRef]
37. Liu, L.; Moon, H.R.; Schorfheide, F. Forecasting with dynamic panel data models. *Econometrica* **2020**, *88*, 171–201. [CrossRef]
38. Pesaran, M.H.; Pick, A.; Timmermann, A. Forecasting with panel data: Estimation uncertainty versus parameter heterogeneity. *Quant. Econ.* **2026**, *17*, 342–393. [CrossRef]
39. Timmermann, A. Forecast combinations. In *Handbook of Economic Forecasting*; Elliott, G., Granger, C.W.J., Timmermann, A., Eds.; Elsevier: Amsterdam, The Netherlands, 2006; Volume 1, pp. 135–196.
40. Claeskens, G.; Magnus, J.R.; Vasnev, A.L.; Wang, W. The forecast combination puzzle: A simple theoretical explanation. *Int. J. Forecast.* **2016**, *32*, 754–762. [CrossRef]
41. Wang, X.; Hyndman, R.J.; Li, F.; Kang, Y. Forecast combinations: An over 50-year review. *Int. J. Forecast.* **2023**, *39*, 1518–1547. [CrossRef]
42. Teräsvirta, T. Forecasting economic variables with nonlinear models. In *Handbook of Economic Forecasting*; Elliott, G., Granger, C.W.J., Timmermann, A., Eds.; Elsevier: Amsterdam, The Netherlands, 2006; Volume 1, pp. 413–457.
43. Hewamalage, H.; Bergmeir, C.; Bandara, K. Recurrent neural networks for time series forecasting: Current status and future directions. *Int. J. Forecast.* **2021**, *37*, 388–427. [CrossRef]
44. Zeng, A.; Chen, M.; Zhang, L.; Xu, Q. Are transformers effective for time series forecasting? *Proc. AAAI Conf. Artif. Intell.* **2023**, *37*, 11121–11128. [CrossRef]
45. Medeiros, M.C.; Vasconcelos, G.F.R.; Veiga, Á.; Zilberman, E. Forecasting inflation in a data-rich environment: The benefits of machine learning methods. *J. Bus. Econ. Stat.* **2021**, *39*, 98–119. [CrossRef]

46. Smyl, S. A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. *Int. J. Forecast.* **2020**, *36*, 75–85. [CrossRef]
47. Pawlowsky-Glahn, V.; Egozcue, J.J.; Tolosana-Delgado, R. *Modeling and Analysis of Compositional Data*; Wiley: Chichester, UK, 2015.
48. Kynčlová, P.; Filzmoser, P.; Hron, K. Modeling compositional time series with vector autoregressive models. *J. Forecast.* **2015**, *34*, 303–314. [CrossRef]
49. Snyder, R.D.; Ord, J.K.; Koehler, A.B.; McLaren, K.R.; Beaumont, A.N. Forecasting compositional time series: A state space approach. *Int. J. Forecast.* **2017**, *33*, 502–512. [CrossRef]
50. Katz, J.N.; King, G. A statistical model for multiparty electoral data. *Am. Political Sci. Rev.* **1999**, *93*, 15–32. [CrossRef]
51. Zheng, T.; Chen, R. Dirichlet ARMA models for compositional time series. *J. Multivar. Anal.* **2017**, *158*, 31–46. [CrossRef]
52. Bergeron-Boucher, M.-P.; Canudas-Romo, V.; Oeppen, J.; Vaupel, J.W. Coherent forecasts of mortality with compositional data analysis. *Demogr. Res.* **2017**, *37*, 527–566. [CrossRef]
53. Egozcue, J.J.; Pawlowsky-Glahn, V.; Mateu-Figueras, G.; Barceló-Vidal, C. Isometric logratio transformations for compositional data analysis. *Math. Geol.* **2003**, *35*, 279–300. [CrossRef]
54. Hyndman, R.J.; Ahmed, R.A.; Athanasopoulos, G.; Shang, H.L. Optimal combination forecasts for hierarchical time series. *Comput. Stat. Data Anal.* **2011**, *55*, 2579–2589. [CrossRef]
55. Wickramasuriya, S.L.; Athanasopoulos, G.; Hyndman, R.J. Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. *J. Am. Stat. Assoc.* **2019**, *114*, 804–819. [CrossRef]
56. Athanasopoulos, G.; Hyndman, R.J.; Kourentzes, N.; Panagiotelis, A. Forecast reconciliation: A review. *Int. J. Forecast.* **2024**, *40*, 430–456. [CrossRef]
57. Giacomini, R.; White, H. Tests of conditional predictive ability. *Econometrica* **2006**, *74*, 1545–1578. [CrossRef]
58. Elliott, G.; Timmermann, A. *Economic Forecasting*; Princeton University Press: Princeton, NJ, USA, 2016.

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.