Multidimensional Panel data regression model: The case of multidimensional home ownership vacancy rate in USA

Talha Omer[1,b], Daniel J Henderson[b], Andros Kourtellos[c]

[a]*Department of Economics, Finance and Statistics, JIBS, Jönköping University, , Jönköping, 55318, Sweden*
[b]*Department of Economics, Finance and Legal studies, Culverhouse Business School, Tuscaloosa, AL 35487, Alabama, United States Of America*
[c]*Faculty of Economics and Management, University of Cyprus , Nicosia, CY 2109 , Cyprus*

Abstract

We expanded the use of structured machine learning in regression for nowcasting using multidimensional panel data that has mixed-frequency series. Our main motivation was the challenge of predicting home ownership vacancy rates across various states and MSAs, especially when the key economic data is sampled at mixed frequencies. We extended the two dimensional panel data [1] sparse group LASSO regularization into multidimensional panel data regression model and it can take advantage of mixed frequency time series multidimensional panel data structure. We successfully employed our proposed extended multidimensional machine learning panel data model to forecast the three dimensional home ownership vacancy rates of USA. The results suggests that our extended multidimensional time series regression model is very useful to nowcast/forecast the home ownership vacancy rates and perform better compare to the traditional time series regression model. Our results are general and our extended multidimensional time series regression model could be applied on any multidimensional macroeconomic problem.

Keywords: Multidimensional panel data, MIDAS, home ownership vacancy rate, sg-LASSO, regularized regression

1. Introduction

Nowcasting is typically a problem of mixed frequency data, for instance, the variable of interest is a low frequency observed series say quarterly, whereas real time information daily, weekly and monthly, during the given quarter can be utilised to assess and nowcast the low frequency series. Conventionally we use the dynamic model to nowcast the object of interest with high frequency data. These dynamic model belongs to state space models and inference can be performed by using the conventional methods ($for\ instance\ Kalman\ filter,\ see$, [2]).Things can get more complicated when we are operating the data rich environments sometime within panel data, since modern research is increasingly relying and adopting the habit of large data set, and this has given the rise to use the multi-dimensional panel data methods in real life empirical studies. Panel data is common in various fields such as social sciences, economics and econometrics. This data consists of multiple individual for instance homeowner ship rate in different states of US over time ($three-dimensional\ panel$), sector level trade between different countries or region ($three-dimensional\ panel$) and so on.This paper extends the two-dimensional panel data model technique from [1] to accommodate multidimensional panel data with mixed-frequency observations.
In context of home ownership rate, US census bureau measure the home ownership rate which is a quarterly data within the states and metropolitan statistical areas (MSA) over different time periods which there are many predictors. From, practically point of view, dealing with the large number of data set it would be hard to handle this with state space models. In the current study, we extend the study of [1] into multidimensional panel data modelling. In such models, time-invariant MSA home ownership specific effects are conventionally used to obtain cross-sectional heterogeneity in data.We would combined this with the regularised regression methods, which is very popular and useful in Economics, and Finance as a useful method to model the predictive relationship via variable selection method. We emphasize on the three dimensional panel data regression models in high dimensional setting where in some situation number of predictors could be larger than the sample size.
To the extent of our knowledge, there are no studies that how to nowcast the three dimensional panel data modeling in case of high-dimensional mixed frequency panels. However, [3], consider the mixed frequency panel data model and did not consider the high dimensional setting while nowcasting and fore-

casting the low frequency series. In a similar manner, [4], implemented mixed frequency VAR panel data model to nowcast the low frequency series but not in high dimensional data setting. Furthermore, [5] develop the spare-group LASSO ($sg - LASSO$) regularize machine learning methods for heavy tailed panel data regression models in context of mixed frequency data and derive the oracle inequalities for the pooled and fixed effects models. However, [1] explores the [5] sg-LASSO model in terms of rich data environment of high dimensional panel setting on panel data, potentially sampled at different frequencies. In this paper, we explore and extend the two dimensional panel data in case of high dimensional into multidimensional panel data in context of high dimensional setting.
We just included the small part of empirical study in this article, and we focus on the quarterly home ownership rate of 75 largest MSA^s of different states of United states. This means we are evaluating the model based approach within quarter prediction for possible shortest horizon. The literature largely emphasize the home ownership rate is a good economic indicator and closely tied to several key aspect of economy, which includes consumer spending, household wealth and health of housing market. Therefore, nowcasting or forecasting the home ownership rate can give the better planning time and help to policy makers, businessman and investor to invest in a efficient way. Stating differently, nowcasting of home ownership rate before the data released can provide the valuable insight to housing market, and broader economy, which can lead to better planning, market efficiency and economic stability.

The paper is organized as follows, Section 2 introduces the multidimensional mixed frequency panel data models and estimators. Section 3 illustrates the simulation design and section 4 reports the out of sample prediction of our extended multidimensional panel data model. Whereas, section 5 represents the results of our empirical application nowcasting of the home ownership rate of different MSA's within states of US. Concluding remarks are given in section 6 Furthermore, all technical and data details is presented in appendix.

2. Multi-dimensional mixed frequency Panel data models

The Panel data regression model can be extended to the choice of dimension depending on the nature of the problem to explore. This paper

motivates the situation when the number of cross-sectional dimensions is large in the panel data setting, and we allow the model set of independent variables to be sampled at m times higher frequency than the dependent variable. Let dependent the variable is observed along three indices, such as y_{igt}, where $i = 1, \ldots .., N_1, g = 1, \ldots, N_2$, and $t = 1, \ldots, 2, T$ and the observation have the same ordering, for example, t is the fastest, then g is the second fastest whereas, the i is the slowest one. For instance, the vector of the dependent variable of three-dimensional panel data can be written as, $(y_{111}, y_{112}, \ldots., y_{11T}, \ldots., y_{1N_21}, \ldots., y_{1N_2T}, \ldots, y_{N_111}, \ldots., y_{N_11T}, \ldots., y_{N_1N_21}, \ldots. y_{N_1N_2T})'$.
Let K be the time-varying independent variables $x_{i,g,(t-(j-1))/(m,k)}$ such that $i \in [N_1], g \in [N_2], t \in [T], j \in [m], k \in [K]$, and we allow the independent variables to be measured at m time higher frequency compared to the dependent variable frequency period $t \in [T]$, every entity $i \in [N_1]$ and $g \in [N_2]$. We consider the following three-dimensional mixed frequency panel data regression model.

$$y_{i,g,t+h} = \upsilon_i + \alpha_i + \gamma_g + \eta_t + \sum_{k=1}^{K} \psi\left(L^{\frac{1}{m}}; \beta_k\right) x_{i,g,t,k} + \epsilon_{i,g,t} \tag{1}$$

where $h \geq 0$ is the prediction horizon, υ_i is the entity specific interest, α_i, and γ_g is the parameter to quantify the individual effect, and η_t is the the time-specific fixed effect, and

$$\psi\left(L^{\frac{1}{m}}; \beta_k\right) x_{i,g,t,k} = \frac{1}{m} \sum_{j=1}^{m} \beta_{j,k} x_{i,g,\frac{t-(j-1)}{m,k}} \tag{2}$$

where $\beta_k = (\beta_{1,k}, \ldots, \beta_{m,k})' \in \mathbb{R}^m$ is a high frequency lag polynomial. If the frequency is specific for each predictor lag $k \in [K]$, then we use m_k instead of m. If $m = 1$, then Equation (1) becomes a standard three-dimensional panel data regression model, which can be written as,

$$y_{igt+h} = \upsilon_i + \alpha_i + \gamma_g + \eta_t + \sum_{k=1}^{K} \beta_k x_{i,g,t,k} + \epsilon_{i,g,t} \tag{3}$$

while $m > 1$ implies that we have m times more observation in the high frequency and then the high-frequency lags of independent variables $x_{i,g,t,k}$ are also included. Realistically, a large number of predictors K with a large number of m (high-frequency measurement) can be an important and rich

source of out-of-sample predictive information. This will lead to laborious and costly jobs to estimate the $N_1 + N_2 + (m \times K)$ parameters and can result in condensed predictive performance in the small sample. Estimation of large number of parameters $N_1 + N_2 + m \times K$ could lead parameter lag proliferation($overfitting$). To overcome the parameters proliferation, we use the MIDAS literature (see, extensive literature,[6];[7];[8];[1];[5]). The literature says that instead of m separate slopes of high-frequency covariate $k\epsilon[K]$ in equation (1) with some constraint on notations, we estimate a weight function ω parameterized by $\beta_k \ \epsilon \ R^L$, where L should be less than m, and equation (2) can be estimated as

$$\psi(L^{1/m}; \beta_k)x_{i,g,t,k} = \frac{1}{m}\sum_{j=1}^{m}\omega(\frac{j-1}{m}; \beta_k)x_{i,g,\frac{t-(j-1)}{m,k}}$$

where weight function ω can be defined as

$$\omega(s; \beta_k) = \sum_{l=0}^{L-1}\beta_{l,k}wl(s) \qquad\qquad \forall_s[0,1]$$

and $(wl)_{l\geq 0}$ is a compilation of L approximating functions, usually, it is called a dictionary in the literature of machine learning. For instance, we can call exponential Almon lag ([9]), Beta polynomial lag specification, infinite polynomial, and many more (see,[8]) for more details. For instance, it could be a fixed set of Legendre polynomials between $[0, 1]$, and it could be calculated by using the Rodrigues formula ($wl(s) = \frac{1}{l!}\frac{d^l}{d_s^l}(s^2 - s)^l$). The first five elements of the Legendre polynomial are $w_1(s) = 1$;
$w_2(s) = 2s - 1$;
$w_3(s) = 6s^2 - 6s + 1$;
$w_4(s) = 20s^3 - 30s^2 + 12s - 1$;
$w_5(s) = 70s^4 - 140s^3 - 90s^2 - 20s + 1$
Every weight function has its own properties, the orthogonal polynomial does have superior numerical properties than the normal non-orthogonal weight function which we discussed earlier. We can use wavelets, trigonometric polynomials, or Gegenbauer polynomials. One of the attractive features of these polynomials is linear in the parameters which could be solved via a convex optimization. We define, $x_i = (X_{ig,1}W, \ldots.., X_{ig,K}W)$,where for each $k \ \epsilon \ [K]$, $X_{ig,k} = (x_{i,g,\frac{t(j-1)}{m,k}})\ j \ \epsilon[m], t \ \epsilon \ [T]$, is a $T \times m$ matrix of predictors and $W = (w_l((j-1)/m)/m)\ j \ \epsilon \ [m], 0 \leq l \leq L-1$ is an $m \times L$ matrix related to

the dictionary $(w_l)l \geq 0$. Then we define our regression equations with the following step:

1. **Variables:** Rather than x_i and y_i, we now have x_{ig} and y_{ig} respectively, signifying the observations for state i within group g.
2. **Time Series Stacking:** For every state i in group g in our example:
$$y_{igt} = l\ v_{ig} + x_{igt}\beta + \epsilon_{igt}$$
where l is an all-ones vector in R^T and $\beta \in R^{LK}$ is a slope coefficients vector.
3. **Covariates:** We redefine x_{ig} to account for the new dimension:
$$x_{ig} = (X_{ig,1}W, \ldots, X_{ig,K}W)$$
For every $k \in [K]$, $X_{ig,k}$ is the matrix of covariates specific to state i in group g. The dictionary W remains constant across all and groups.
4. **Regression Equation:** After stacking the time series observations for each i in group g:
$$y_{ig} = l\alpha_{ig} + x_{ig}\beta + u_{ig}$$
5. **Stacking Cross-sectional Observations:** For the three dimensions, the stacking becomes:
$$y = (y_{11}^T, \ldots, y_{1N_2}^T, \ldots, y_{N_1 1}^T, \ldots, y_{N_1 N_2}^T)^T$$
$$X = (x_{11}^T, \ldots, x_{1N_2}^T, \ldots, x_{N_1 1}^T, \ldots, x_{N_1 N_2}^T)^T$$
$$\epsilon = (\epsilon_{11}^T, \ldots, \epsilon_{1N_2}^T, \ldots, \epsilon_{N_1 1}^T, \ldots, \epsilon_{N_1 N_2}^T)^T$$
The final regression equation is:
$$y = DB + X\beta + \epsilon$$

Where B is the Kronecker product of the identity matrix for firms and groups with the all-ones vector for time, that is, $B = (I_{N_1} \otimes I_{N_2}) \otimes l$.

The vector y can be defined as:

$$y = \begin{bmatrix} y_{11}^T \\ \vdots \\ y_{1N_2}^T \\ \vdots \\ y_{N_1 1}^T \\ \vdots \\ y_{N_1 N_2}^T \end{bmatrix}$$

where B is called the composite fixed effect of the parameter and defined as $B = (\alpha^{'}\gamma^{'}\eta^{'})$ with $\alpha^{'} = (\alpha_1, ..., \alpha_{N_1})$, $\gamma^{'} = (\gamma_1, ..., \gamma_{N_2})$, $\eta^{'} = (\eta_1, ..., \eta_T)$, $D = ((I_{N_1} \otimes l_{N_2}T), (l_{N_1} \otimes I_{N_2} \otimes l_T), (\ l_{N_1}l_{N_2} \otimes I_T)\ )$ and the last term ϵ is the disturbance term. The MIDAS approach not only benefits to accommodate the mixed frequency covariates but also successfully help us to reduce the dimensionality concern to the high frequency lags. In case of a small mismatch in the frequencies of the dependent and independent variables, the Unrestricted MIDAS (UMIDAS) scheme developed by[10] could be efficient and able to estimate the coefficient parameters connected with each high-frequency covariates separately. The detailed derivation of the UMIDAS scheme can be seen in [10], and for the three-dimensional panel data regression model, this approach may not be very attractive in high dimensional setting and more than one high-frequency regressors due to overfitting problems in case of a large number of estimations of the coefficients (see., for details, [1],[10]).

When there are a large number of potential predictors K, utilizing extra regularization techniques can enhance the predictive accuracy in situations where the sample size is small. We would take the advantage of sg-LASSO regularization of [5] and extend it to sg-LASSO regularization to three dimensional panel data regression model. The fixed effect sg-LASSO estimator $\hat{\rho} = \left(\hat{\alpha}^{\mathrm{T}}, \hat{\beta}^{\mathrm{T}}\right)^{\mathrm{T}}$ gives

$$\min_{a,b \in \mathbb{R}^{N+p}} |y - DB - Xb|^2_{N_1 \times N_2 \times T} + 2\lambda\Omega(b) \tag{4}$$

where Ω is the sg-LASSO regularizing function. We do not include the intercept in the design matrix X, and it means we don't penalize the fixed effects which we assume not sparse. The empirical norm is defined as, $|\cdot|^2_{N_1 \times N_2 \times T} = |\cdot|^2/(N_1 \times N_2 \times T)$ and $\Omega(b) = \gamma|b|1 + (1-\gamma)|b|2, 1$ is regularization function (see details, [1]).

Similarly, the pooled sg-LASSO estimator $\hat{\rho} = \left(\hat{\alpha}^{\mathrm{T}}, \hat{\beta}^{\mathrm{T}}\right)^{\mathrm{T}}$ for three dimensional panel data given as

$$\min_{r=(a,b) \in \mathbb{R}^{1+p}} |y - l\upsilon_i + x_i b|^2_{N_1 \times N_2 \times T} + 2\lambda\Omega(r) \tag{5}$$

In case of large $N_1 \times N_2 \times T$ Pooled regression considers appealing, but can result in lost of heterogeneity of individual time series. The next section is devoted to the simulation design and later the results of simulation design, which is the extension of [5] into multidimensional panel data setting.

3. Simulation design for three dimensional MIDAS panel data model

To evaluate the predictive performance of three-dimensional panel data models, we simulate the data from the following two data-generating processes (DGP). In the first DGP, we will generate the data for $K = 6$ regressors whereas, the second DGP would be more about in a high dimensional setting like [1], we include the $K = 24$ monthly regressors and ten noisy covariates.

$$y_{ig,t} = \nu + \rho_1 y_{ig,t-1} + \frac{1}{m}\sum_{j=1}^{m} \omega(\frac{j-1}{m}; \beta_k) x_{i,g,\frac{t-(j-1)}{m,k}} + \epsilon_{igt} \qquad \text{(DGP1)}$$

$$\begin{aligned} y_{ig,t} = \nu + \rho_1 y_{ig,t-1} + \rho_1 y_{ig,t-2} + \alpha_i + \gamma_g + \eta_t \\ + \sum_{k=1}^{K} \frac{1}{m}\sum_{j=1}^{m} \omega(\frac{j-1}{m}; \beta_k) x_{i,g,\frac{t-(j-1)}{m,k}} + \epsilon_{igt} \end{aligned} \qquad \text{(DGP2)}$$

where $i \in [N_1]$, $g \in [N_2]$, $t \in [T]$, ν is the common intercept, $\frac{1}{m}\sum_{j=1}^{m} \omega(\frac{j-1}{m}; \beta_k)$ $x_{i,g,\frac{t-(j-1)}{m,k}}$the weight function for the k_{th} high frequency covariate, and the error terms $\epsilon_{igt} \sim_{i.i.d} N(0,1)$. The target variable of interest $y_{ig,t}$ is driven by one auto regressive lag and two auto regressive lags augmented with high-frequency variables for DGP_1 and DGP_2, respectively, and thus it is a pooled MIDAS three-dimensional panel data model. Furthermore, $\rho_1 = 0.5$ and $\rho_2 = 0.05$, and the high frequency regressor $K = 1$ and $K = 6$,for (DGP_1) and (DGP_2), respectively. Our primary interest is quarterly/monthly (dependent/regressor) mixed data, which will give us four lags for each high-frequency regressors for each time period.

3.1. ***Monte carlo experiments***

We investigate the finite sample out of sample prediction of the methods so far. We consider the unstructured UMIDAS [10] elastic net and sg-LASSO with MIDAS. We have utilized the two tuning parameters λ and γ. The γ parameter defined as the relative weight of of LASSO, ridge and group LASSO. Whereas, the γ will interpolate the between LASSO and ridge in elastic net UMIDAS. In both cases we report the results in three different grid values $\gamma \in \{0, 0.5, 1\}$. Furthermore, we select the tuning parameter λ after the selection of three values of γ. We consider the $K - fold$ cross-validation criteria for the three dimensional panel data setting and creates

fold based on cross sectional units instead of pooled sample. We consider the 5-fold cross validation for our simulation experiment. We further consider two information criteria to evaluate our extended approach. We consider the following two performance criteria: AIC [11],[12] and BIC [13]. We assume that $y_{i,g\ ,t}$ given that $x_{i,g\ ,t}$ are drawn independently and identically from normal distribution and the likelihood can be written as;

$$L(v,\beta,\sigma^2) \propto -\frac{1}{2\sigma^2} N_1 \times N_2 \sum_{i=1}^{N_1}\sum_{g=1}^{N_2}\sum_{t=1}^{T}(y_{i,g,t} - v_i - \mathbf{x}_i^{\top}\beta)^2$$

The AIC criteria can be written as

$$AIC = \frac{\|y - \hat{\epsilon} - X\hat{\beta}\|^2_{N_1\times N_2\times T}}{N_1 \times N_2 \times T \times \hat{\sigma}^2} + \frac{2}{N_1 \times N_2 \times T} \times \hat{df}$$

and the BIC criterion can be written as

$$BIC = \frac{\|y - \hat{\epsilon} - X\hat{\beta}\|^2_{N_1\times N_2\times T}}{N_1 \times N_2 \times T \times \hat{\sigma}^2} + \frac{\log(N_1 \times N_2 \times T)}{N_1 \times N_2 \times T} \times df,$$

where df stands for degree of freedom, $\hat{\epsilon} = \hat{v}l$ and $\hat{\epsilon} = \hat{v}D$, for pooled regression and fixed effect regression. For more details, ([5]; [1]; [14]). We certainly believe that AIC suppose to perform well in case we have large K is large compare to the sample size.

4. Monte Carlo simulation results

Table 1 reports the average mean squared error of out-of-sample prediction. We report the results of pooled panel data and fixed-effect estimators in the left block and right block in Table 1, respectively. Overall reported results appear to be in line with [5] two-dimensional panel data, however, we consider the multidimensional panel data settings. The sg-LASSO-MIDAS performs better compared to Unstructured elnet UMIDAS and average sg-MIDAS in all DGP's settings. In the case of sg-LASSO-MIDAS, the best performance is achieved for $\gamma\ \epsilon\ \{0,1\}$ for both pooled panel data and fixed effect case, whereas, when $\gamma = 0$ LASSO solution seems to dominate in the case of elastic net UMIDAS for both the pooled and fixed effect cases. However, when $\gamma = 1$,in a group LASSO, a substantial improvement in prediction quality is observed when the MIDAS polynomial is compared with

the UMIDAS and average MIDAS.

	$N_1 \times N_2 \times T$	Pooled panel data			Fixed effects		
		$\gamma = 0$	0.5	1	$\gamma = 0$	0.5	1
AIC							
sg-LASSO	$5 \times 20 \times 48$	4.439	4.410	4.444	4.451	4.453	4.446
	$10 \times 20 \times 48$	4.151	4.161	4.171	4.117	4.162	4.641
	$10 \times 20 \times 60$	4.171	4.141	4.161	4.123	4.132	4.117
Unrestricted elnet MIDAS	$5 \times 20 \times 48$	16.652	16.632	16.622	16.672	16.682	16.952
	$10 \times 20 \times 48$	4.368	4.358	4.378	4.368	4.368	4.368
	$10 \times 20 \times 60$	4.368	4.368	4.151	4.181	4.191	4.201
Average sg-MIDAS	$5 \times 20 \times 48$	15.372	15.321	15.381	15.371	15.391	15.395
	$10 \times 20 \times 48$	4.298	4.291	4.288	4.258	4.268	4.308
	$10 \times 20 \times 60$	4.310	4.321	4.341	4.331	4.231	4.251
BIC							
sg-LASSO	$5 \times 20 \times 48$	4.205	4.211	4.251	4.261	4.31	4.34
	$10 \times 20 \times 48$	4.251	4.251	4.251	4.251	4.251	4.251
	$10 \times 20 \times 60$	4.351	4.355	4.361	4.371	4.359	4.362
Unrestricted elnet MIDAS	$5 \times 20 \times 48$	15.951	15.982	15.971	15.981	15.991	15.998
	$10 \times 20 \times 48$	4.447	4.456	4.466	4.731	4.746	4.846
	$10 \times 20 \times 60$	4.368	4.547	4.568	4.668	4.478	4.488
Average sg-MIDAS	$5 \times 20 \times 48$	16.736	16.736	16.736	16.736	16.736	16.736
	$10 \times 20 \times 48$	3.876	3.876	3.876	3.876	3.876	3.876
	$10 \times 20 \times 60$	4.298	4.298	4.298	4.298	4.298	4.298
Cross validation							
sg-LASSO	$5 \times 20 \times 48$	4.131	4.111	4.141	4.161	4.128	4.191
	$10 \times 20 \times 48$	4.151	4.161	4.171	4.181	4.181	4.21
	$10 \times 20 \times 60$	4.16	4.191	4.201	4.231	4.251	4.261
Unrestricted elnet MIDAS	$5 \times 20 \times 48$	4.41	4.478	4.481	4.486	4.478	4.491
	$10 \times 20 \times 48$	4.568	4.568	4.561	4.668	4.671	4.888
	$10 \times 20 \times 60$	4.668	4.723	4.91	4.912	4.991	4.993
Average sg-MIDAS	$5 \times 20 \times 48$	4.937	4.972	4.972	4.981	4.982	4.997
	$10 \times 20 \times 48$	4.698	4.719	4.981	4.988	4.991	5.121
	$10 \times 20 \times 60$	4.998	4.898	4.798	4.898	4.998	4.981

Table 1: The table reports simulation results for out-of-sample prediction accuracy for fixed effect estimators and for the DGP_1 for the sg-LASSO-MIDAS and elastic net unrestricted MIDAS. We vary the cross-sectional dimensions $N_1 \epsilon \{5, 10\}$ and $N_2 \epsilon \{20\}$ and time series dimensions $T \epsilon \{48, 60\}$. We report results for 5-fold cross-validation, AIC and BIC information criteria λ tuning parameter calculation methods and for three different values 0, 0.5 *and* 1 of γ.

In the case of pooled panel data, increasing N_1 from 5 to 10 seems to have a larger impact on the performance compared to an increase in the time series dimension 48 to 60. This is because increasing the time series dimensions

give us more parameters to estimate. However, in the case of fixed effect results, the difference remains somehow sharper than the pooled panel data. We have presented the DGP_2 results in the appendix Table 5. In the higher dimensions setting, the sg-LASSO dominates good compared to the other methods. However, the performance of the UMIDAS become worst in high dimensional setting. In the end, when comparing the results among different selection methods, that are two different information criteria and cross-validation. From Table 1, we observe that in all situations cross-validation leads to a smaller prediction error in both pooled panel data and fixed effect panel data cases. However, changing the value of γ does not play a significant role, and this shows that our results are very much consistent with the result of [1]. By comparing the AIC with BIC, BIC remains the worst compared to AIC especially when we combined it with the Unrestricted MIDAS and small N_1,N_2, and time dimensions values.

5. Nowcasting of home ownership vacancy rate

We extended the two dimensional panel data nowcasting approach of [5] and [1] to multidimensional panel. However, [15] and [1] documented that analyst could make a systematic and predictable error while ignoring the mixed frequencies data set. In this section, we therefore consider to nowcast home ownership vacancy rate of 75 largest MSA's of USA within State using a set of predictor which are typically sampled at different frequencies. We use 5 predictors including traditional economic indicator. We applied the pooled and fixed effects sg-LASSO MIDAS model, and compare our extended approach with several benchmark models, which includes, random walk(RW), and elastic net UMIDAS model.
We also compute the prediction of two dimensional panel, using state i and time T, and similarly, MSA g and time T. We will present the prediction results for multidimensional panel sg-LASSO MIDAS model and compare it with two dimensional panel sg-LASSO MIDAS model. This comparison aims to highlight the advantages and enhanced the functionality of multidimensional panel when applied in a high dimensional context, assuming that the adequate data is available. The remaining of the section is as follows, we present the description of our data and more detailed of results, followed by a summary of method and empirical methods.

5.1. Data description

The home ownership vacancy rate in USA considered to be very crucial key factor of economic stability and social mobility. Despite being leading economic indicator, home ownership vacancy rate could not get enough attention in the literature. Home ownership vacancy rate, as per its definition, refers to the proportion of the privately owned home in specific region or country that are currently occupied and available for sale. Historically, United States is the combination of states and states are further divided with different metropolitan statistical areas ($MSA's$) and micropolitan statistical areas. Home ownership vacancy rate is the quarterly index, and measure by US census bureau by ($https://www.census.gov/$). Home ownership vacancy rate varies state to state and also within MSA of these states [16]. Home ownership vacancy rate could be lower in the larger MSA's due to variety of factors. One could be the high cost of housing in these areas, which make it harder for the people to save money to pay the mortgage and down payment. We collect the home ownership vacancy rate quarterly data of 75 largest MSA's of the US within the 39 states of US, from the US census bureau. The full sample consists of observations between the 1^st of March, 2005 and the 30^th of December, 2022. Due to lagged dependent variables in the models, our effective sample starts from third fiscal quarter of 2005. We use the 32 observations for the initial sample and remaining 42 observations for evaluating out-of- sample forecast for each MSA, which we obtain by using the expanding window forecasting scheme. Our target variable is home ownership vacancy rate of each 75 largest MSA's of 39 different states of US. Home ownership vacancy rate data are subject to delay between 1 to 2 months, for instance, the first quarter data of home ownership vacancy rate is released on 3^rd May, 2023. We used the same data which we got from the website. We will provide the list of the states and MSA at the end of the this paper. Our analysis commences with the evaluation of the stationarity of our variables. There appears to be a dearth of literature addressing stationarity checks for multidimensional panel data. Therefore, we undertake the task by analyzing the data in a couple of distinct way. Initially, we treat the two dimensions namely, g and t, as the top 75 largest MSA's of USA and the their corresponding time periods for each variables. Subsequently, we average out the State i over MSA and t as second dimension. Since we don't have homogeneous panel, we run the cross sectional panel dependent unit root test for heterogeneous panel [17] and [18]. Upon completion of the analysis, we employed the CIP test to assess the stationarity, and we observe

the absence the of the unit root in our data.

5.2. Models and main results

We estimate the several regression model to compute the forecast. We begin to estimate with the individual sg-LASSO MIDAS regression for the each MSA $i = 1, \cdots, N$, and we refer the model as individual,

$$y_i = lv_i + x_i\beta_i + \epsilon_i$$

where MSA predictions are computed at $\hat{y}_{i,t+1} = \hat{v}_i + \mathbf{x}_{i,t+1}^{\top}\hat{\beta}_i$. We noted in the 2, x_i contains the lags of target variable which is low frequency variable and high frequency variable, so we apply the Legendre polynomials of degree 3, as applied by [1]. The next step we estimate the two dimensional panel data pooled and fixed effect sg-LASSO MIDAS regression data models

$$y = lv_i + X\beta + \epsilon \quad \text{(Pooled)}$$

$$y = D\alpha + X\beta + \epsilon \quad \text{(Fixed Effects)}$$

and similarly we can compute the predictions of above given models as,

$$\hat{y}_{i,t+1} = l\hat{v}_i + \mathbf{x}_{i,t+1}^{\top}\hat{\beta} \quad \text{(Pooled)}$$

$$\hat{y}_{i,t+1} = \hat{v}_i + \mathbf{x}_{i,t+1}^{\top}\hat{\beta} \quad \text{(Fixed Effects)}$$

at the next we estimate our extend multidimensional pooled and fixed effects sg-LASSO MIDAS regression panel data models,

$$\hat{y}_{i,t+1} = l\hat{v}_i + \mathbf{x}_{i,t+1}^{\top}\hat{\beta} \quad \text{(Pooled)}$$

$$\hat{y}_{i,t+1} = \hat{v}_i + \mathbf{x}_{i,t+1}^{\top}\hat{\beta} \quad \text{(Fixed Effects)}$$

We benchmark the home ownership vacancy rate and panel data regression against two simple alternatives. First, we compute forecasts for the RW model as

$$\hat{y}_{i,t+1} = y_{i,t}$$

Second we consider predictions of home ownership vacancy rate implied by home ownership vacancy rate nowcasts using the information up to time $t+1$, and it can be written as,

$$\hat{y}_{i,t+1} = \bar{y}_{i,t+1} \tag{6}$$

where $\hat{y}$ is considered the forecasted home ownership vacancy rate, which is made at the end of $t+1$ quarter. We already mentioned that the real home ownership vacancy rates are available and delayed by two months at the end of quarter. We measure the performance of our extended method by considering the mean squared forecast error (MSFE) for all methods. And the general expressions of the MSFE are,

$$\text{MSFE} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T-T_{is}+1}(\bar{y}_i - \hat{y}_i)^\top(\bar{y}_i - \hat{y}_i)$$

where, $\hat{y}_i = (y_{i,T_{Ls}+1}, \ldots, y_{i,T_{Fs}})^\top$ out of sample home ownership vacancy rate values, where T_{Ls} and T_{Fs} represents the last in-sample observation for the first prediction and the last out-of-sample observations, respectively. Whereas, $\hat{\mathbf{y}}_i = (\hat{y}_{i,t_{Ls}+1}, \ldots, \hat{y}_{i,t_{Os}})$ denote the out-of-sample forecast.
At first step, by utilizing aggregated data, we calculate the present quarter's nowcast for the home ownership vacancy rate. This is done for both two-dimensional panel data, represented by state (i) and time (T), and three-dimensional panel data, depicted as state (i), MSA (g), and time (T) respectively. Out of sample RMSE of the both models are presented in the Table 2. Both models performs good in either way, but in terms of prediction accuracy three dimensional panel regression perform much better and gained much more prediction accuracy because of the additional MSA information.We are considering two different models to determine whether multidimensional panel data regression enhances the prediction accuracy. However, we have limited literature on the forecasting of home ownership vacancy rates, for instance, [19]; forecasted the home ownership vacancy rates based on urban areas, and [20] consider the three different states and time as panel data to see the home ownership rate external benefits which is closely related to home ownership vacancy rate. However, performing a three dimensional panel data regression model is very complex task, but literature also says that it would reward in terms of prediction accuracy [21].

Panel data regression models based on aggregation data

	Fixed effect	Pooled	Individual
2D, State i, Time T	2.081	1.07	1.637
3D, State i, MSA g, and Time T	1.295	1.280	1.394

Table 2: The table presents the root mean square error (RMSE) of out-of-sample predictions for the current quarter's home ownership vacancy rates in the United States, derived from aggregated data.

At second step, we comeback to our original claim that MIDAS regression works better when the mixed frequency data is present. We considered the overall mixed frequency time series data of home ownership vacancy rate and run the sg-LASSO, Unrestricted elnet LASSO and average LASSO on mixed frequency data to compute the predictive accuracy of the individual MSA home ownership vacancy rate. We nowcast current quarter ahead home ownership vacancy rate, and report the out of sample absolute RMSE in Table 3. The reported results in suggests that the overall performance of the regularized regression method models improves the prediction accuracy. The sg-LASSO performs better and gain the prediction accuracy in our empirical application, regardless of changing the tuning parameter γ, as it does not make any significant impact as already observed in [1]. Panel data regression regression models performs better due to additional cross sectional information, but aggregation of the mixed frequency data could potentially lose the important information and it can reduce the prediction accuracy [6],[7].
In our study, we implemented two-dimensional and three-dimensional panel data regression models on aggregated data to compare their performance. However, we found that the regularized MIDAS regression, which leverages mixed frequency data without necessitating aggregation, displayed superior performance when run on the same overall time series regression data. The main point is that aggregating data can result in losing important information. The regularized MIDAS regression model effectively addresses this issue. As a result, we observed a notable improvement in prediction.

	γ=0	γ=0.5	γ=1
sg-LASSO	1.31	1.30	1.30
Average sg-MIDAS	1.33	1.33	1.33
Unrestricted elnet MIDAS	1.31	1.32	1.31

Table 3: Table reports the out-of-sample home ownership vacancy rates for overall time series regression data based on mixed frequency data by using the different regularized regression method by using three different tuning parameter values 0, 0.5 *and* 1 of γ.

Consequently, we took our research a step further, applying the regularized regression approach to two-dimensional panel data. And the results are reported in Table 4.Turning to the comparison of model based predictions, we see from the results in Table 4 , sg-LASSO MIDAS panel data improves the quality of predictions in comparison to the average sg-LASSO and unrestricted elnet MIDAS regardless of tuning parameter. This suggest that panel data structure is relevant to nowcast the home ownership vacancy rate. Among the panel data, we see that pooled regression in most cases improves the prediction results compare to the fixed effect regression.

Panel: States and Time

	Pooled			Fixed effect		
	γ=0	γ=0.5	γ=1	γ=0	γ=0.5	γ=1
	Cross validation					
sg-LASSO	1.34	1.33	1.32	1.59	1.59	1.58
Average sg-MIDAS	1.26	1.27	1.26	1.57	1.55	1.55
Unrestricted elnet MIDAS	1.32	1.36	1.36	1.61	1.62	1.63
	AIC					
sg-LASSO	1.31	1.33	1.32	1.59	1.58	1.59
Average sg-MIDAS	1.29	1.30	1.29	1.57	1.57	1.56
Unrestricted elnet MIDAS	1.32	1.34	1.36	1.58	1.59	1.60
	BIC					
sg-LASSO	1.36	1.36	1.31	1.56	1.56	1.58
Average sg-MIDAS	1.26	1.26	1.26	1.53	1.55	1.55
Unrestricted elnet MIDAS	1.29	1.29	1.31	1.23	1.58	1.59

Panel A: States, MSA and Time

	Pooled			Fixed effect		
	γ=0	γ=0.5	γ=1	γ=0	γ=0.5	γ=1
	Cross validation					
sg-LASSO	1.15	1.14	1.16	1.21	1.20	1.19
Average sg-MIDAS	1.23	1.22	1.20	1.25	1.26	1.29
Unrestricted elnet MIDAS	1.29	1.30	1.30	1.32	1.36	1.39
	AIC					
sg-LASSO	1.21	1.13	1.30	1.29	1.24	1.21
Average sg-MIDAS	1.27	1.27	1.26	1.32	1.32	1.31
Unrestricted elnet MIDAS	1.35	1.36	1.34	1.35	1.40	1.39
	BIC					
sg-LASSO	1.13	1.20	1.18	1.15	1.16	1.17
Average sg-MIDAS	1.34	1.34	1.34	1.46	1.46	1.45
Unrestricted elnet MIDAS	1.38	1.36	1.39	1.41	1.41	1.42
Random walk RMSE	1.44					

Table 4: The table reports home ownership vacancy rate out-of-sample prediction accuracy for fixed effect, and pooled effect estimators and for two different panels by using the sg-LASSO-MIDAS average MIDAS, and elastic net unrestricted MIDAS. We report results for 5-fold cross-validation, AIC and BIC information criteria λ tuning parameter calculation methods and for three different values 0, 0.5 *and* 1 of γ.

Eventually, this led us to our extension - a three-dimensional regularized panel data regression model, which showed even greater potential in handling complex datasets as evident by our findings presented in Table 2, the incorporation of MSA as a third dimension resulted in an additional enhancement of prediction accuracy, even when based on aggregated data. This progression in our research signifies the crucial role of comprehensive, multi-dimensional data analysis in yielding more accurate predictions and stronger model performance. We reported the absolute RMSE of the one step ahead nowcast of home ownership vacancy rate 4. We observed the our extended three dimensional panel data model from different angle. First of all, it performs better than the traditional multidimensional panel data regression model, which is considered to be the high dimensional panel data regression.Second within the regularized machine learning approaches sg-LASSO estimator performs better in terms of prediction accuracy with cross validation and BIC, and results are inline with the finding of [1] . Overall, the unrestricted elnet MIDAS remains shaky in the process in the process and performed worst compare to the average MIDAS and sg-LASSO estimator,however, in some cases it performs better as compare to the aggregative multidimensional panel data regression, even from the random walk forecast as well. Unrestricted MIDAS could be shaky and parameters over-fitting could appear, especially when we do have more than one predictor [10], so purpose was to test the unrestricted MIDAS combine with machine learning methods. In our study, sg-LASSO estimator which we extended for the three dimensional panel data regression model well suited for incorporating the grouped fixed effect. This approach involves grouping state specific effect intercept based on either statistical procedures or economics reasoning, as similar the literature outlined by [22]. Our presented results in Table 4, suggests that the use of group fixed effect improves the accuracy of our nowcast. Considerably, when we check with the best choice of tuning parameter choice, group fixed effects outer performs other panle data models, especially with BIC criteria. Therefore, our findings suggest that grouped fixed effect performs better between capturing heterogeneity and pooled parameters resulting in more accurate nowcast predictions. We performs the Diebold and Mariano (DM) test[23] to compare the predictive accuracy between models. Since, we have so many models to compare we just report some of the models in the appendix in Table 6. Overall, all the models performs better when $\gamma = 1$ which is a group LASSO solution. Overall summary of the results, give an early nowcast of the home ownership vacancy yields a better prediction accuracy compare to

the analyst prediction in terms of prediction accuracy. In addition, these findings highlights the potential benefit of leveraging the machine learning techniques, especially when we do have high dimensional setting, complexity of estimation of large number of parameters that could lead to the less accurate prediction.

6. Conclusion

Home ownership vacancy rates plays a vital role in as an economics indicator especially in the real estate, especially in US context where data is generally collected at the national, state and MSA level. However, home ownership vacancy rates is measured quarterly and very much dependent on differnt high frequency covariates, like unemployment rates. Traditionally, research employ the two dimensional panel data regression based on states and time as primary dimension. In this article, we build upon on the methodology introduced by [1] and [5] namely, sg-LASSO, elnet unrestricted MIDAS, and average MIDAS, and extended the model into multidimensional one. Simulations using two distinct DGPs, followed by nowcasting the home ownership vacnacy rates allowed us to incorporate the state, MSA and time as dimensional an analytical work. We have presented our results from broader perspective. First, by looking at home ownership vacancy rates on three dimensions states, MSA and time gives more gain in the accuracy of nowcast. At the end, We illustrates that incorporating the regularized multidimensional MIDAS panel data regression performs better compared to the regular multidimensional panel data regression based on aggregated data. Our approach is general and can be employed on any multidimensional panel data , which can eventually give us the timely updates on the important economic indicators, like home ownership vacancy rates.

7. Sample Appendix Section

	$N_1N_2 \times T$	Pooled panel data			Fixed effects		
		$\gamma = 0$	0.5	1	$\gamma = 0$	0.5	1
AIC							
sg-LASSO	$5 \times 20 \times 48$	4.208	4.221	4.231	4.241	4.258	4.256
	$10 \times 20 \times 48$	4.598	4.560	4.567	4.576	4.569	4.701
	$10 \times 20 \times 60$	4.642	4.612	4.599	4.601	4.621	4.612
Unrestricted elnet MIDAS	$5 \times 20 \times 48$	20.663	20.671	20.672	20.683	20.691	20.701
	$10 \times 20 \times 48$	4.105	4.121	4.141	4.151	4.161	4.171
	$10 \times 20 \times 60$	4.161	4.171	4.181	4.195	4.201	4.213
Average sg-MIDAS	$5 \times 20 \times 48$	13.763	13.775	13.791	13.791	13.797	13.811
	$10 \times 20 \times 48$	4.014	4.012	4.201	4.341	4.621	4.721
	$10 \times 20 \times 60$	4.780	4.791	4.810	4.723	4.741	4.731
BIC							
sg-LASSO	$5 \times 20 \times 48$	3.795	3.801	3.811	3.821	3.831	3.841
	$10 \times 20 \times 48$	4.212	4.271	4.261	4.281	4.291	4.293
	$10 \times 20 \times 60$	4.598	4.601	4.560	4.561	4.621	4.631
Unrestricted elnet MIDAS	$5 \times 20 \times 48$	20.775	20.781	20.791	20.795	20.799	20.231
	$10 \times 20 \times 48$	4.377	4.382	4.384	4.386	4.394	4.395
	$10 \times 20 \times 60$	4.101	4.241	4.231	4.211	4.201	4.191
Average sg-MIDAS	$5 \times 20 \times 48$	15.353	15.363	15.373	15.383	15.393	15.396
	$10 \times 20 \times 48$	4.151	4.142	4.158	4.149	4.151	4.176
	$10 \times 20 \times 60$	4.012	4.131	4.231	4.245	4.261	4.246
Cross validation							
sg-LASSO	$5 \times 20 \times 48$	4.208	4.208	4.208	4.208	4.208	4.208
	$10 \times 20 \times 48$	4.151	4.161	4.171	4.181	4.181	4.21
	$10 \times 20 \times 60$	4.16	4.191	4.201	4.231	4.251	4.261
Unrestricted elnet MIDAS	$5 \times 20 \times 48$	4.264	4.257	4.257	4.264	4.257	4.257
	$10 \times 20 \times 48$	4.568	4.568	4.561	4.668	4.671	4.888
	$10 \times 20 \times 60$	4.668	4.723	4.91	4.912	4.991	4.993
Average sg-MIDAS	$5 \times 20 \times 48$	3.685	3.685	3.685	3.685	3.685	3.685
	$10 \times 20 \times 48$	4.698	4.719	4.981	4.988	4.991	5.121
	$10 \times 20 \times 60$	4.998	4.898	4.798	4.898	4.998	4.981

Table 5: The table reports simulation results for out-of-sample prediction accuracy for fixed effect estimators and for the DGP_2 for the sg-LASSO-MIDAS and elastic net unrestricted MIDAS. We vary the cross-sectional dimensions $N_1 \epsilon \{5, 10\}$ and $N_2 \epsilon \{20\}$ and time series dimensions $T \epsilon \{48, 60\}$. We report results for 5-fold cross-validation, AIC and BIC information criteria λ tuning parameter calculation methods and for three different values 0, 0.5 *and* 1 of γ.

Comparison	DM Statistic	p-value
AIC, $\gamma = 1$, fe, vs Average, AIC, $\gamma = 1$	-2.582	0.004997
BIC, $\gamma = 1$, pool, vs Average, AIC, $\gamma = 1$	-6.0443	1.136e-09
BIC, $\gamma = 1$, fe, cv vs Average, $\gamma = 1$, pool	-5.8168	4.291e-09
AIC, $\gamma = 1$, fe vs Average, $\gamma = 1$, fe	-1.9678	0.02472
BIC, $\gamma = 1$, pool vs Average, $\gamma = 1$, fix	-4.7027	1.504e-06
BIC, $\gamma = 1$, fe vs Average AIC, $\gamma = 1$, fe	-4.9373	4.795e-07
BIC, $\gamma = 1$, pool vs Average, pool, $\gamma = 1$	-3.9062	5.073e-05
BIC, $\gamma = 1$, pool vs Average, pool, $\gamma = 1$	-3.8068	7.56e-05

Table 6: Summary of Standout Diebold-Mariano Test Results

References

[1] A. Babii, R. T. Ball, E. Ghysels, J. Striaukas, Machine learning panel data regressions with heavy-tailed dependent data: Theory and application, Journal of Econometrics (2022).

[2] M. Bańbura, D. Giannone, M. Modugno, L. Reichlin, Now-casting and the real-time data flow, in: Handbook of economic forecasting, Vol. 2, Elsevier, 2013, pp. 195–237.

[3] L. Khalaf, M. Kichian, C. J. Saunders, M. Voia, Dynamic panels with midas covariates: nonlinearity, estimation and fit, Journal of Econometrics 220 (2) (2021) 589–605.

[4] J. Fosten, R. Greenaway-McGrevy, Panel data nowcasting, Econometric Reviews 41 (7) (2022) 675–696.

[5] A. Babii, R. T. Ball, E. Ghysels, J. Striaukas, Panel data nowcasting in a data-rich environment: The case of price-earnings ratios, Available at SSRN (2022).

[6] E. Ghysels, P. Santa&Clara, R. Valkanov, Vthe midas touch: Mixed data sampling regression models, V UNC and UCLA Working Papers (2002).

[7] E. Ghysels, P. Santa-Clara, R. Valkanov, The midas touch: Mixed data sampling regression models (2004).

[8] E. Ghysels, A. Sinko, R. Valkanov, Midas regressions: Further results and new directions, Econometric reviews 26 (1) (2007) 53–90.

[9] S. Almon, The distributed lag between capital appropriations and expenditures, Econometrica: Journal of the Econometric Society (1965) 178–196.

[10] C. Foroni, M. Marcellino, C. Schumacher, Unrestricted mixed data sampling (midas): Midas regressions with unrestricted lag polynomials, Journal of the Royal Statistical Society. Series A (Statistics in Society) (2015) 57–82.

[11] H. Akaike, Statistical predictor identification, Annals of the institute of Statistical Mathematics 22 (1) (1970) 203–217.

[12] H. Akaike, Information theory and an extension of the maximum likelihood principle,[w:] proceedings of the 2nd international symposium on information, bn petrow, f, Czaki, Akademiai Kiado, Budapest (1973).

[13] H. Akaike, A bayesian analysis of the minimum aic procedure, arın, fyrst, Statist, Marh 3 (1978) 90–14.

[14] H. Zou, T. Hastie, R. Tibshirani, On the "degrees of freedom" of the lasso (2007).

[15] R. T. Ball, E. Ghysels, Automated earnings forecasts: Beat analysts or combine and conquer?, Management Science 64 (10) (2018) 4936–4952.

[16] K. C. Bishop, N. V. Kuminoff, N. A. D. Murphy, Tax policy and the heterogeneous costs of homeownership (2023).

[17] K. S. Im, M. H. Pesaran, Y. Shin, Testing for unit roots in heterogeneous panels, Journal of econometrics 115 (1) (2003) 53–74.

[18] M. H. Pesaran, A simple panel unit root test in the presence of cross-section dependence, Journal of applied econometrics 22 (2) (2007) 265–312.

[19] J. Lee, G. Newman, Forecasting urban vacancy dynamics in a shrinking city: A land transformation model, ISPRS International Journal of Geo-Information 6 (4) (2017) 124.

[20] N. E. Coulson, H. Li, Measuring the external benefits of homeownership, Journal of Urban Economics 77 (2013) 57–67.

[21] B. H. Baltagi, Panel data forecasting, Handbook of economic forecasting 2 (2013) 995–1024.

[22] S. Bonhomme, E. Manresa, Grouped patterns of heterogeneity in panel data, Econometrica 83 (3) (2015) 1147–1184.

[23] F. DIEBOLD, R. MARIANO, Comparing predictive accuracy. journal of business and economics statistics, v. 13 (1995).