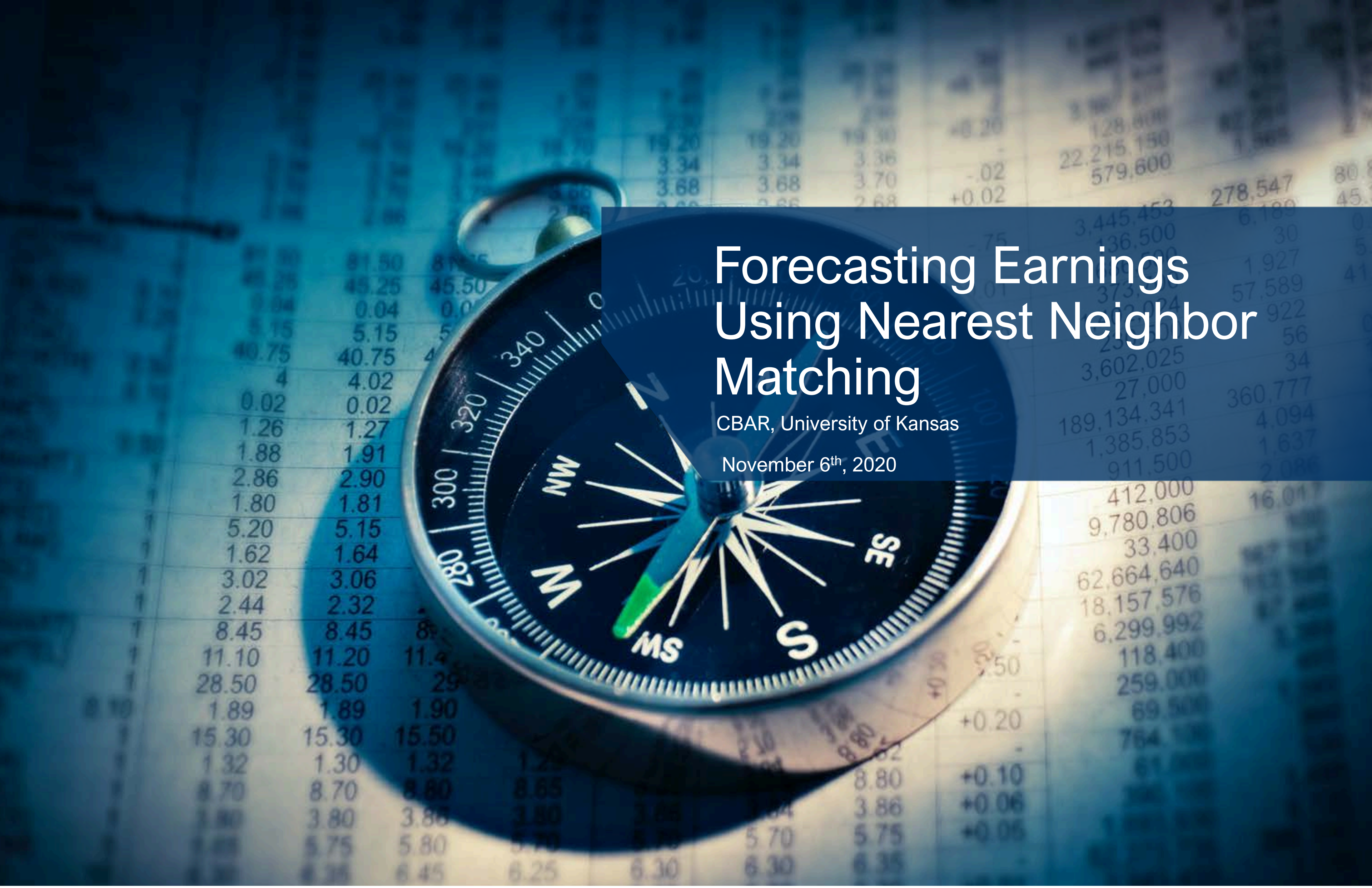


Forecasting Earnings Using Nearest Neighbor Matching

CBAR, University of Kansas

November 6th, 2020

Peter Easton, Martin Kapons, Steven Monahan, Harm Schuett, Eric Weisbrod

Tilburg University

David Eccles School of Business, The University of Utah

University of Notre Dame, Mendoza College of Business

KU School of Business, The University of Kansas

This paper in a nutshell

What?

- Develops a new, simple earnings forecasting method

Why?

- Practical: to provide more accurate earnings forecasts for a large sample of firms
- Theoretical: to learn more about how past and future earnings are linked

How?

- Find and use “comparable” time-series of firms

This is how we teach our class:

ACC 672 List of Project Company Pairs

Air Delivery & Freight	
Federal Express	FDX
United Parcel Service	UPS

Beverages	
Coca-Cola	KO
PepsiCo	PEP

Consulting	
Booz Allen Hamilton	BAH
Accenture	ACH

Processed & Packaged Foods	
General Mills	GIS
Kellogg	K

Farm & Construction Machinery	
Caterpillar	CAT
Deere & Company	DE

This is not a good match:

But what about this?

McDonald's circa 1960s

Chipotle circa 2020

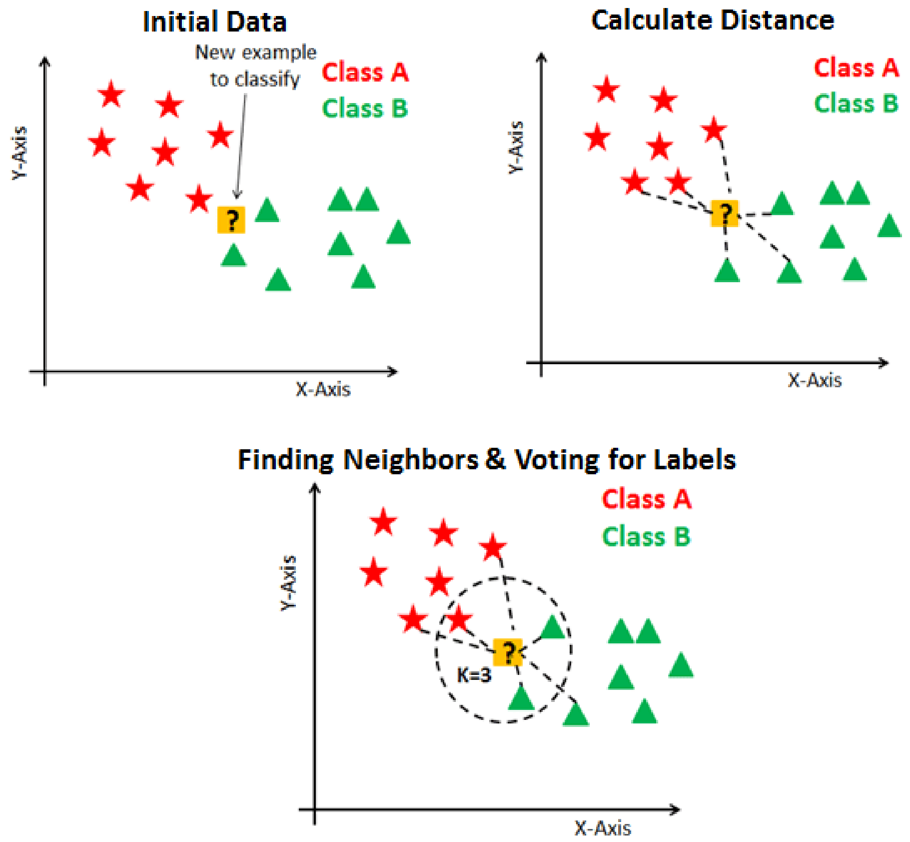
Let the data speak: *K* Nearest Neighbor Approach

- "Things that appear similar are likely similar"

  (Chen & Shah, 2018)

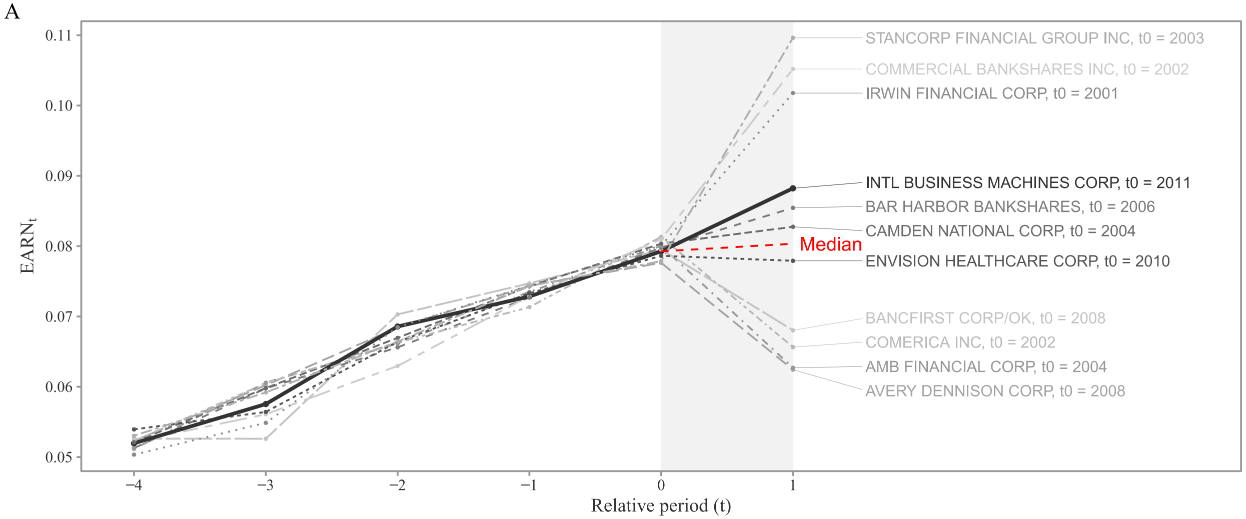
- Method appears as early as 11th century, *Book of Optics*

- Nate Silver examples:
  - Baseball players
  - Similarities of US states

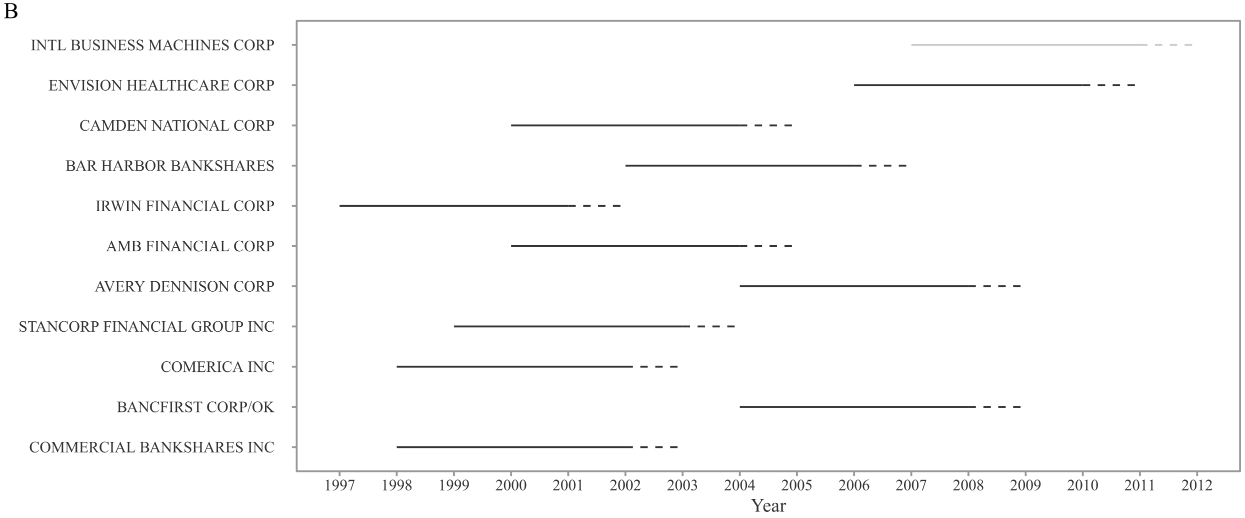


Example - Figure 1: Forecasting IBM 2012 using KNN

A



B



Takeaways

1. Simple Nearest Neighbor Match (NNM) is most accurate
    1. Simple match on 2 years of earnings history
    2. Many peers (~80)

2. NNM more accurate than random-walk or other regression-based forecasts over short and long forecast horizons
3. NNM not statistically worse than analyst forecasts in 3 out 4 forecast error metrics. Performs well in cases where no analyst forecasts are available

→ Significant information in past earnings about future earnings – KNN approach exploits it better than other approaches

Remainder of the presentation: 3 RQs

and why!

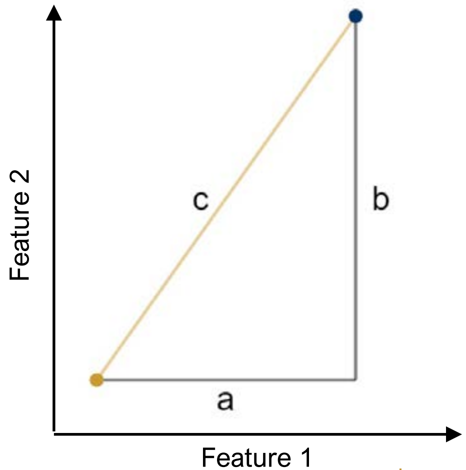
1. How can KNN prediction methods (best) be used to forecast earnings?
2. How accurate are NNM forecasts compared to competing approaches?
3. When are NNM forecasts more or less likely to outperform other approaches?

RQ1: How can KNN prediction methods (best) be used to forecast earnings?

- For each sequence $i, t: t - M + 1$, find the *K* most similar among all possible sequences $j, s: s - M + 1$ during the previous 10 years of panel data
- Most similar is measured via Euclidian distance

$$DIST_{i,t,j,s}^{F,M} = \sqrt{\sum_{f=1}^{F} \sum_{m=1}^{M} \left(FEAT_{i,t-m+1}^{f} - FEAT_{j,s-m+1}^{f}\right)^2}$$

- Potentially important considerations
  - Features *FEAT*
  - Length of sequence *M*
  - Number of sequences *K*

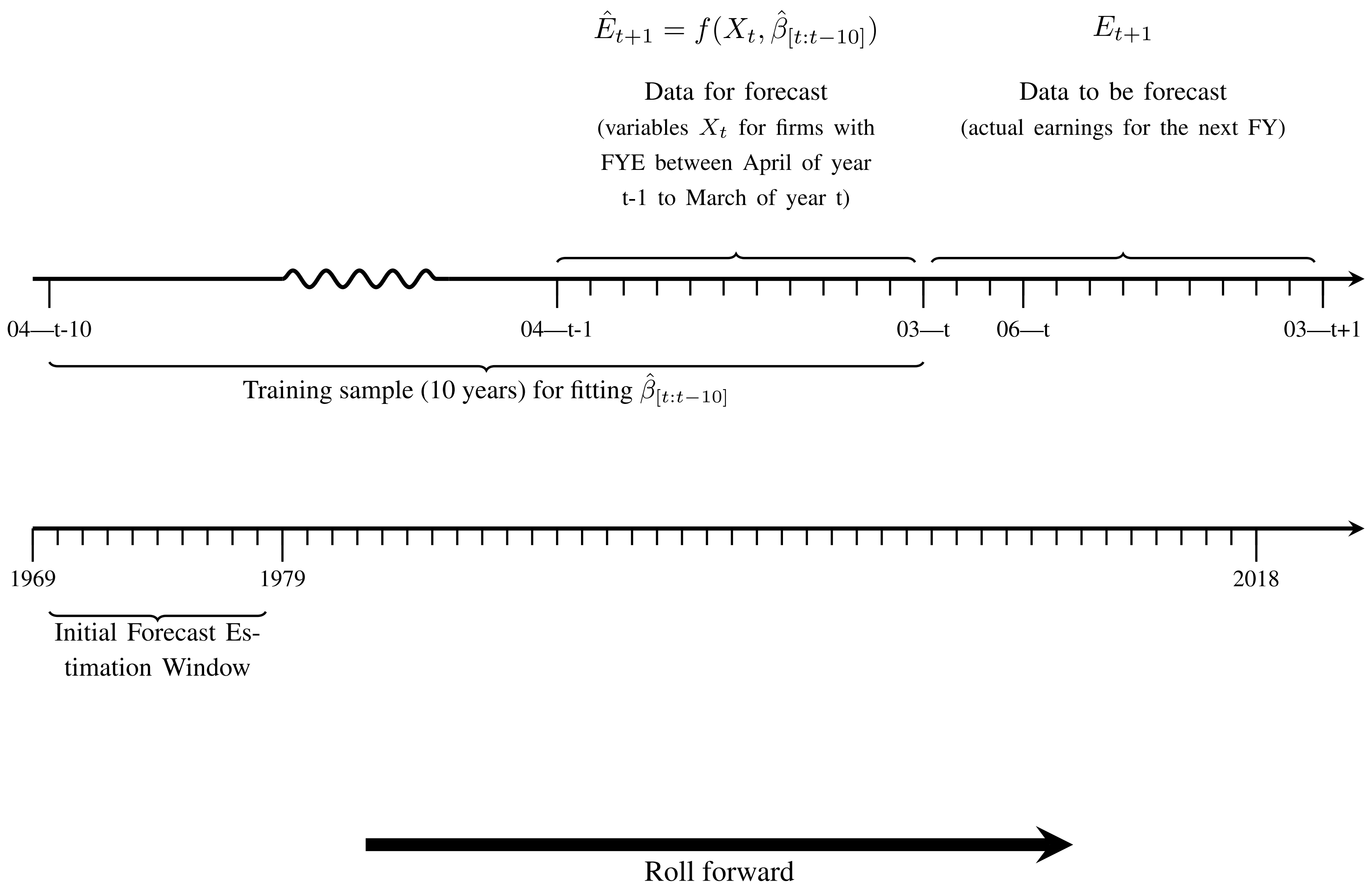


RQ1: How can KNN prediction methods (best) be used to forecast earnings?

- Once the neighbors are identified:
- Forecast firm i's earnings using the future scaled earnings of the median neighbor *j*:

$$\mathbb{E}_t\left[EARN_{i,t+1}\right] = EARN_{j,s+1}$$

Figure 2: Rolling out-of-sample forecasting approach

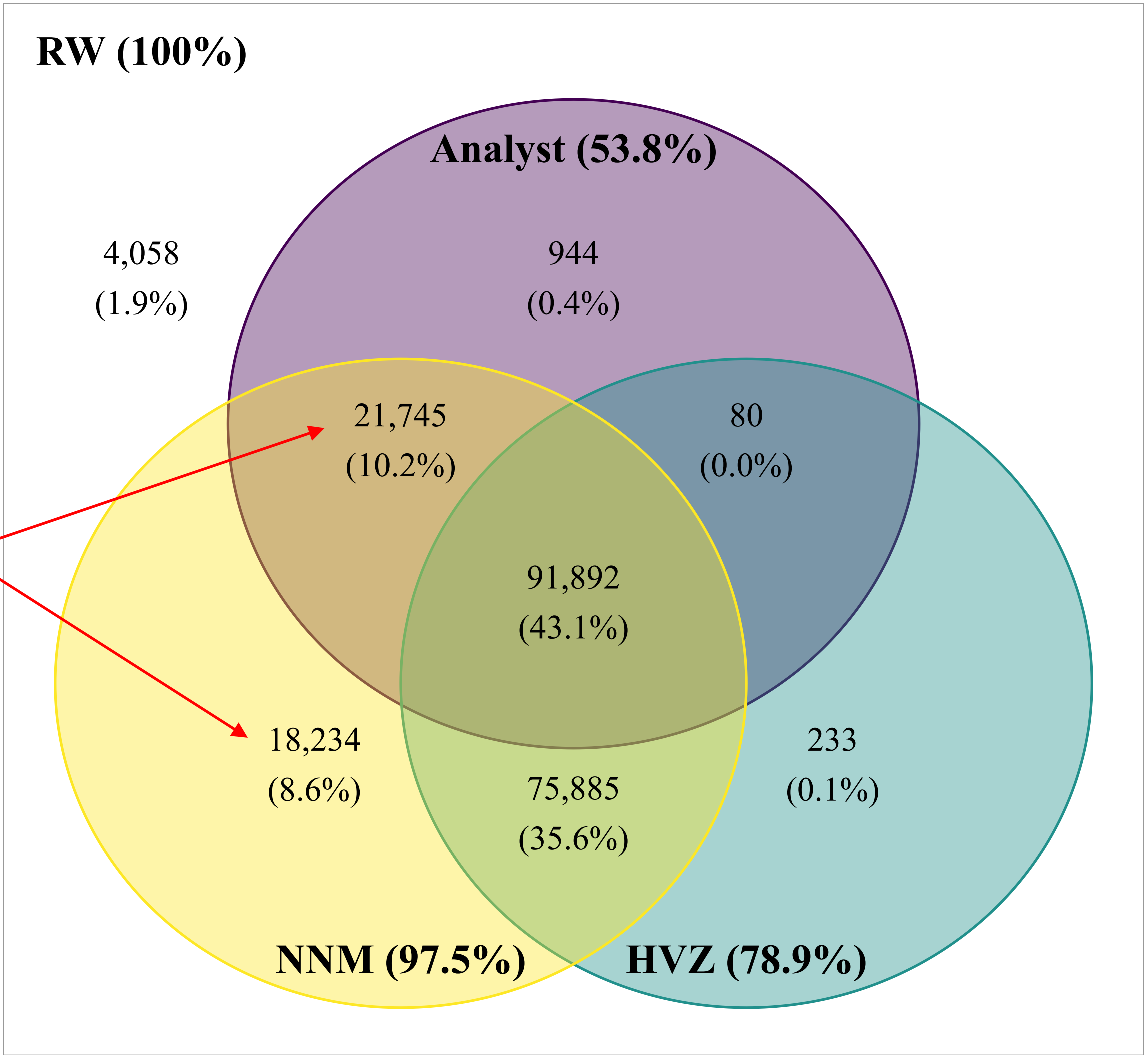


Data and variables

Data Filter	Firm-Years
Total Compustat Observations 1979 – 2018	339,171
Less missing EBSI	-53,214
Less missing and non-positive deflators	-72,886
Random walk forecast sample	213,071
Less missing lagged EBSI	-5,315
Nearest neighbor matching forecast sample	207,756
Less missing accruals	-39,979
Less missing future EBSI	-11,092
Less MVE < $10M	-24,646
Forecast Comparison Sample	132,039

$EBSI_{i,t}$	Earnings before special items for firm i at time t
$MVE_{i,t}$	Equity market value for firm i at the end of fiscal year t
$EARN_{i,t}$	$EBSI_{i,t}$ scaled by $MVE_{i,t}$
$FEARN_{i,t+h}$	Forecast of $EBSI_{i,t+h}$ scaled by $MVE_{i,t}$

Figure 3: Forecast Coverage by Model



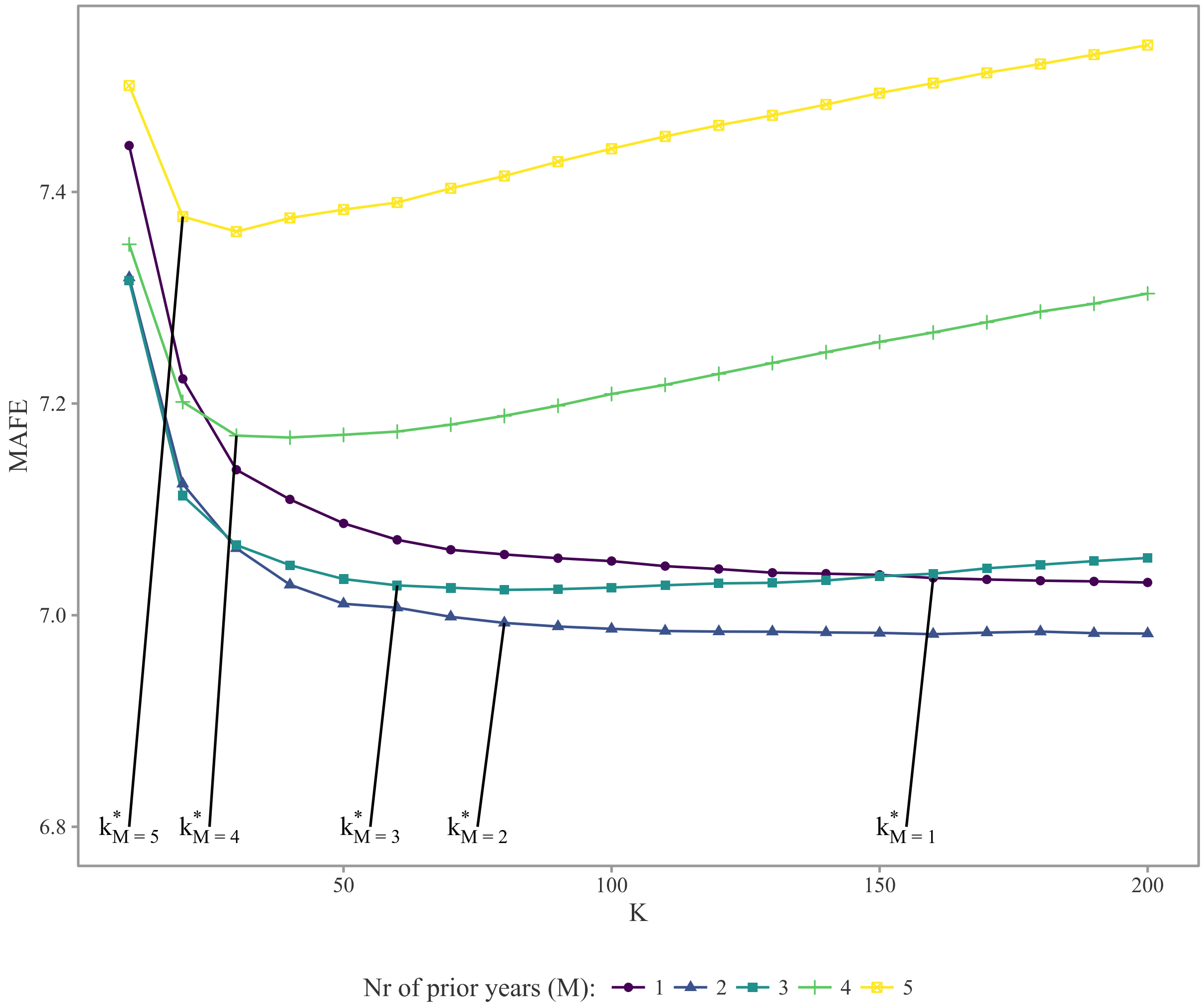
39,979 more forecasts than HVZ

Quantifying forecast accuracy

MAFE	Mean absolute forecast error	Mean($\lvert EBSI_{i,t+h}$ - $FEBSI_{i,t+h} \rvert$ / $MVE_{i,t}$) * 100
MDAFE	Median absolute forecast error	Median($\lvert EBSI_{i,t+h}$ - $FEBSI_{i,t+h} \rvert$ / $MVE_{i,t}$) * 100
MSE	Mean of squared forecast error	Mean((($EBSI_{i,t+h}$ - $FEBSI_{i,t+h}$) / $MVE_{i,t})^2$) * 100
TMSE	Mean of squared forecast error after truncating the top and bottom 0.1% signed forecast errors	

- Computed on the whole time series 1979 – 2018
- Compute accuracy differences using regressions

Best combination of history length (M) and number of sequences (K)



Note:
FEAT = EARN

Different variations of NNM

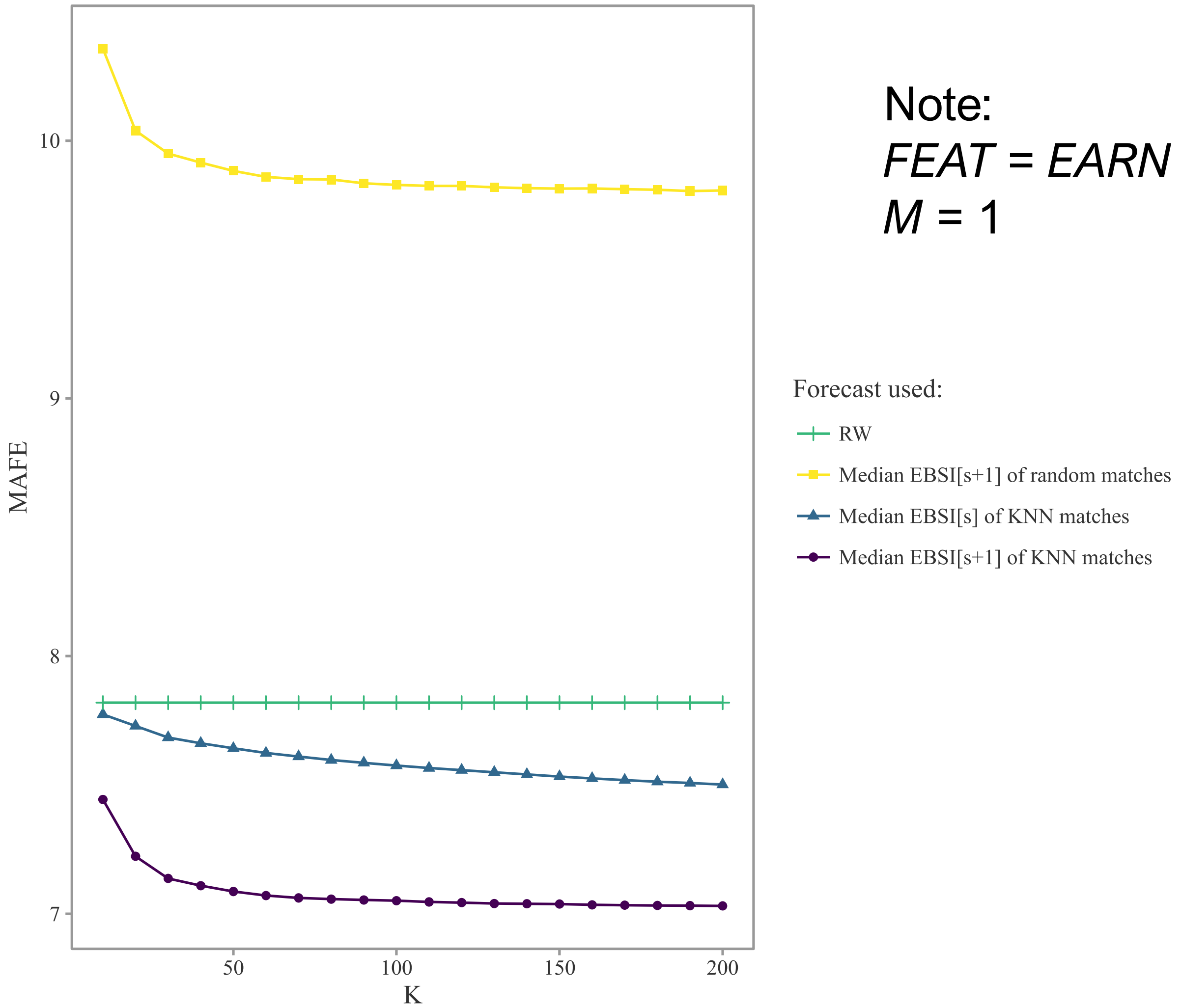
Model	N	MAFE	MDAFE	MSE	TMSE
(a) t+1 forecast error for different deflators					
$NNM1_{MVE}$	166,824	6.965	2.412	7.631	1.973
$NNM1_{BVE}$	166,824	0.330***	0.063***	5.190	0.135***
$NNM1_{TA}$	166,824	0.452***	0.051***	6.675	0.234***
$NNM1_{Sale}$	166,824	0.507***	0.065***	6.034*	0.284***
(b) t+1 forecast error for different matching variables					
NNM1	125,484	6.893	2.529	6.716	1.797
NNM2	125,484	0.076***	0.055***	0.141***	0.067***
NNM3	125,484	0.232***	0.190***	0.210**	0.151***
(c) t+1 forecast error for different stratification clusters					
$NNM1_{MVE}$	166,824	6.965	2.412	7.631	1.973
NNM_{FF12}	166,824	0.074***	0.018	0.405**	0.132***
NNM_{Size}	166,824	0.103***	0.033**	0.549***	0.159***
(d) t+1 forecast error vs BCG model					
$NNM1_{MVE}$	166,270	6.950	2.409	7.614	1.958
BCG	166,270	0.923***	0.210***	5.728**	0.611***
BCG_{MVE}	166,270	0.702***	0.211***	3.368**	0.362***

Note: NNM1 only uses *EARN* as feature. NNM2 adds *ACC*, NNM3 adds full set of HVZ variables

Why does NNM work?

- Wisdom of the crowds
  - Similar to analyst consensus forecast
- Company *i* might have a weird year, median neighbor unlikely to be weird
- Earnings generally grow
  - Current NNM is model weaker for loss firms

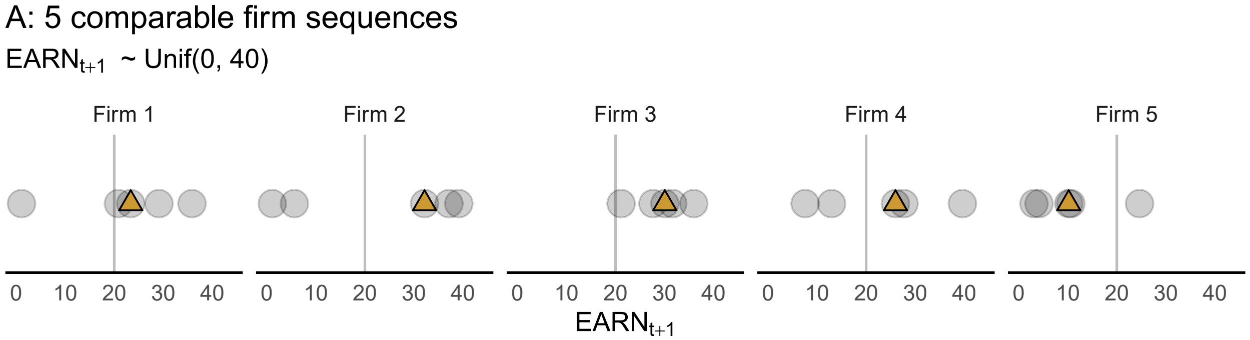
Fig 4: benefits from KNN matching decomposed

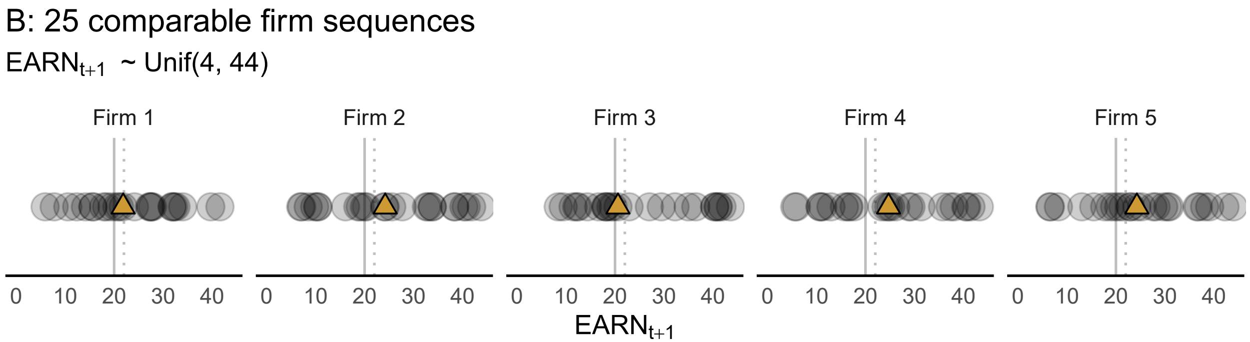


Note:
FEAT = *EARN*
M = 1

The intuition behind large K

Even though you might get less comparable sequences, larger K reduce the variance of the estimator. A simulated example:





Why does use of one feature—earnings—provide the most accurate forecasts?

- “Curse of Dimensionality” – the more variables, the harder it is to find matches
- Earnings and earnings growth (PE and PEG ratios) are the most fundamental indicators of value (Ohlson and Juettner-Nauroth, 2005)

Conclusion RQ1

1. Short history (M = 2) works best
2. Earnings only model works best
3. No stratification by industry, etc. necessary

Short-term earnings sequence alone already contains significant amount of information about future earnings when put into proper context

RQ2: How accurate are nearest NNM forecasts compared to those from competing approaches?

Compare NNM vs RW and HVZ for t+1, t+2, t+3

- NNM:

$$\mathbb{E}_t\left[EARN_{i,t+1}\right] = EARN_{j,s+1}$$

- RW:

$$\mathbb{E}_t[EARN_{t+1}] = EARN_t$$

- Hou, Van Dijk, and Zhang [JAE 2012] (HVZ model):

$$\mathbb{E}_t[EARN_{t+1}]$$
$$= \alpha_0 + \alpha_1 TA_{i,t} + \alpha_2 D_{i,t} + \alpha_3 DIV_{i,t} + \alpha_4 EARN_{i,t} + \alpha_5 LOSS_{i,t}$$
$$+ \alpha_6 ACC_{i,t}$$

Forecast accuracy comparison

Model	N	MAFE	MDAFE	MSE	TMSE
(a) t+1 forecast error					
NNM	132,039	6.876	2.557	6.413	1.769
RW - NNM	132,039	0.727***	0.111***	4.303	0.423***
HVZ – NNM	132,039	2.553***	1.451***	3.617***	1.161***
(b) t+2 forecast error					
NNM	121,097	8.867	3.936	5.448	2.546
RW – NNM	121,097	1.210***	0.117***	7.628**	1.005***
HVZ – NNM	121,097	1.277***	0.591***	4.697**	0.815***
(c) t+3 forecast error					
NNM	110,908	10.531	4.872	7.483	3.461
RW – NNM	110,908	1.284***	0.144***	5.015***	1.340***
HVZ – NNM	110,908	1.422***	0.804***	2.476*	1.020***
(d) (t+1) + (t+2) + (t+3) aggregate forecast error					
NNM	110,666	22.155	10.356	30.252	14.895
RW – NNM	110,666	3.127***	0.299***	34.354***	7.304***
HVZ – NNM	110,666	3.168***	1.638***	17.903***	4.492***

NNM forecast accuracy versus analysts (STREET earnings)

Model	N	MAFE	MDAFE	MSE	TMSE
(a) t+1 forecast error					
NNM	96,345	7.672	1.584	2350.366	1.389
ANALYST	96,345	-0.955	-0.322***	-56.176	0.031
RW	96,345	0.356	0.113***	-40.871	0.532***
(b) t+2 forecast error					
NNM	74,679	6.419	2.336	87.634	1.047
ANALYST	74,679	-0.661	-0.130***	-81.512	0.109***
RW	74,679	0.435	0.210***	-60.223	0.626***

Conclusion RQ2

- NNM uniformly better than other model-based approaches
- Even over long horizons
- “Similar” i.e., not statistically different from analysts in 3 out of 4 metrics

RQ3: When are NNM forecasts more or less likely to outperform other approaches?

- Are there certain situations where NNM does especially well/poor?
- Comparisons by coverage and period
- Cross-sectional splits
- Comparisons by industry
- Can we predict future returns?
- Can we predict ex-ante when NNM does well?

Forecast accuracy by coverage

Model	N	MAFE	MDAFE	MSE	TMSE
(a) t+1 forecast error with analyst coverage					
NNM	86,272	5.482	2.033	5.091	1.145
RW	86,272	0.640***	0.142***	5.027	0.252***
HVZ	86,272	2.055***	1.334***	1.744***	0.688***
(b) t+1 forecast error without analyst coverage					
NNM	45,767	9.505	3.931	8.906	3.114
RW	45,767	0.892***	0.077*	2.937***	0.764***
HVZ	45,767	3.493***	1.739***	7.149***	2.207***

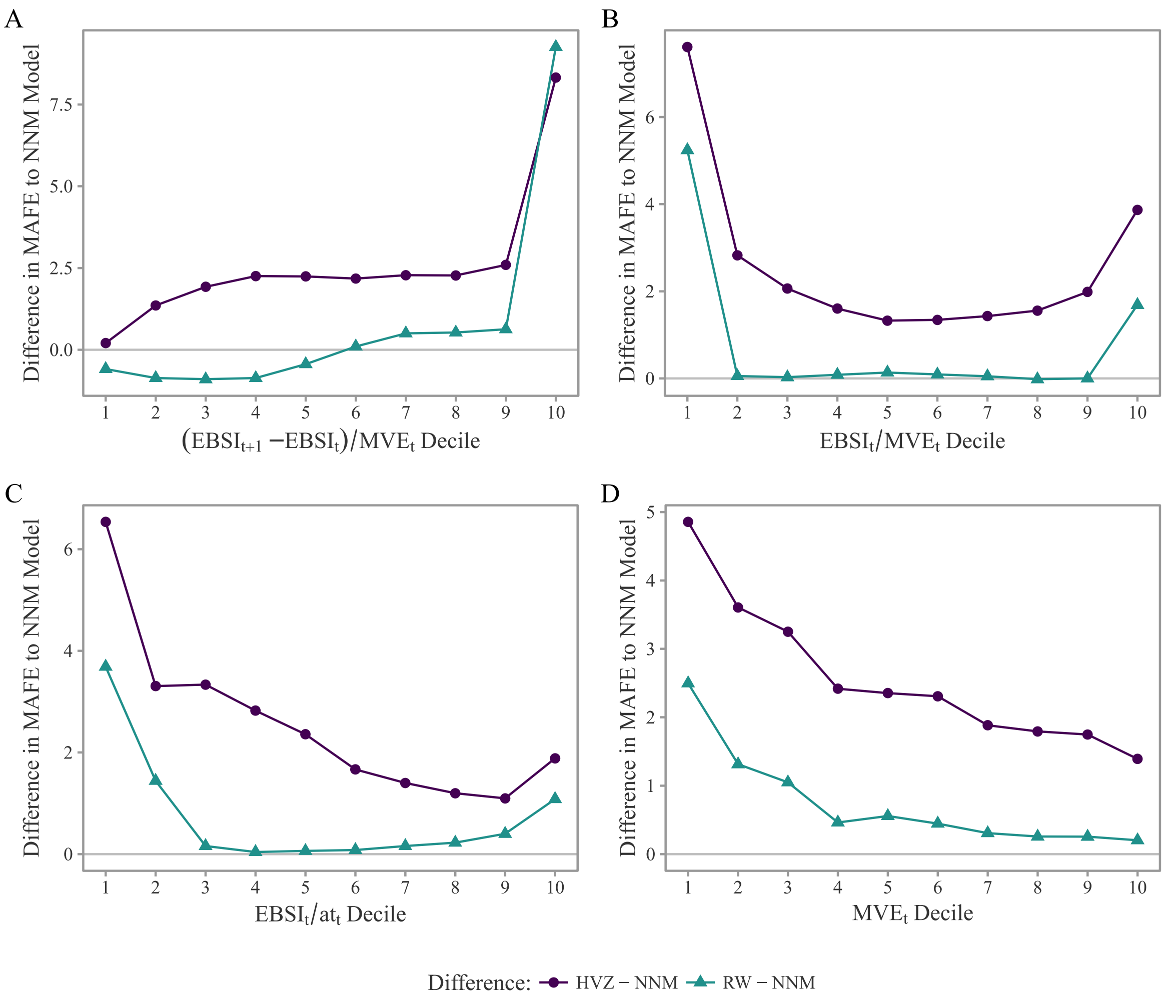
Forecast accuracy by period

Model	N	MAFE	MDAFE	MSE	TMSE
(a) t+1 forecast error, 1979 – 1988					
NNM	28,749	6.375	2.758	2.305	1.455
RW	28,749	0.220	0.156***	0.548	0.020
HVZ	28,749	3.257***	1.505***	4.159**	1.579***
(b) t+1 forecast error, 1989 – 1998					
NNM	37,485	5.799	2.463	1.753	1.139
RW	37,485	0.339***	0.102***	0.266***	0.148***
HVZ	37,485	1.567***	1.003***	0.992***	0.475***
(c) t+1 forecast error, 1999 – 2008					
NNM	37,582	7.675	2.635	11.228	2.199
RW	37,582	1.016***	0.071**	2.995**	0.818***
HVZ	37,582	3.096***	1.683***	6.215***	1.796***
(d) t+1 forecast error, 2009 – 2017					
NNM	28,223	7.755	2.365	10.377	2.763
RW	28,223	1.374***	0.107***	15.230	1.014***
HVZ	28,223	2.423***	1.732***	3.094***	1.331***

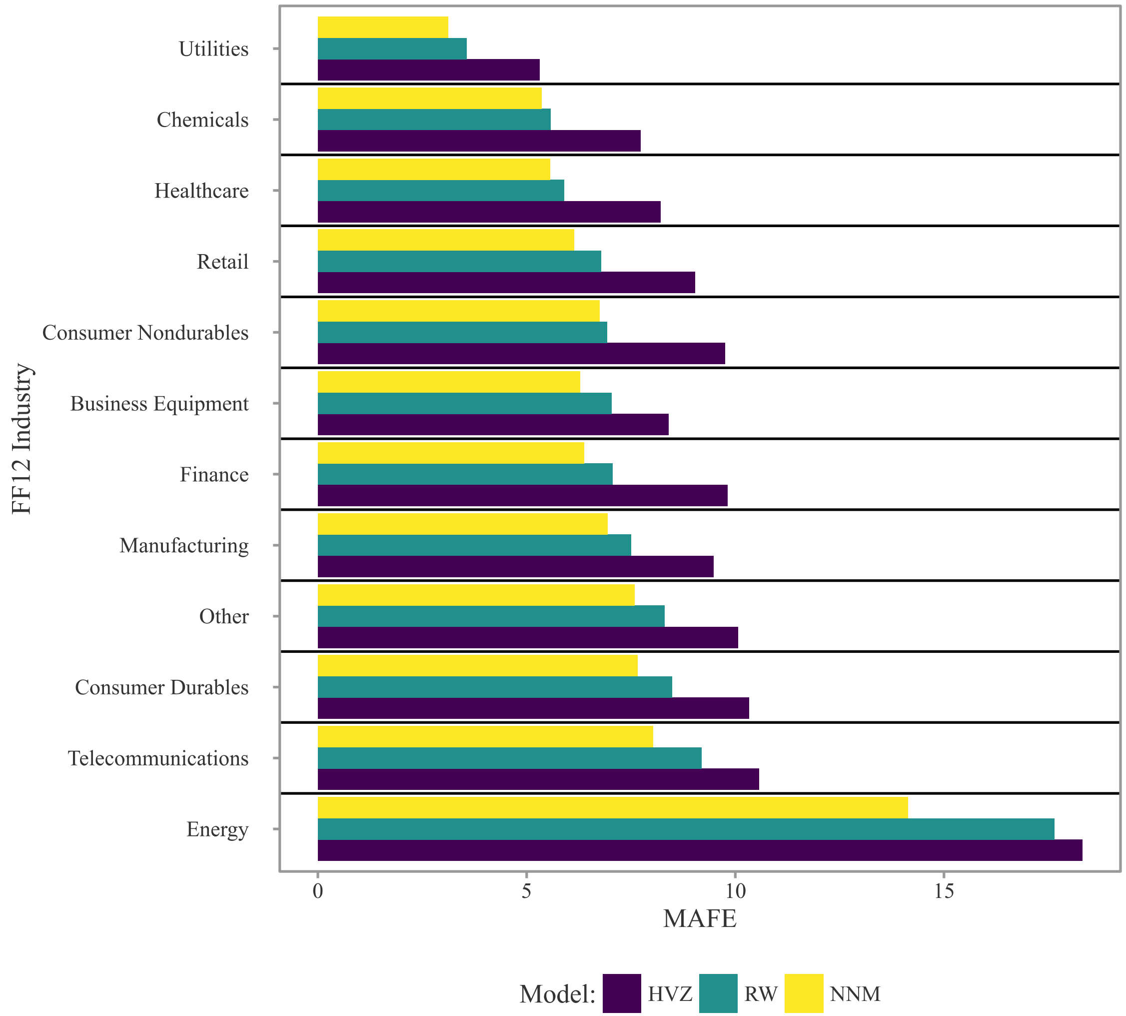
Forecast accuracy by period

Model	N	MAFE	MDAFE	MSE	TMSE
(a) t+1 forecast error, 1979 – 1988					
NNM	28,749	6.375	2.758	2.305	1.455
RW	28,749	0.220	0.156***	0.548	0.020
HVZ	28,749	3.257***	1.505***	4.159**	1.579***
(b) t+1 forecast error, 1989 – 1998					
NNM	37,485	5.799	2.463	1.753	1.139
RW	37,485	0.339***	0.102***	0.266***	0.148***
HVZ	37,485	1.567***	1.003***	0.992***	0.475***
(c) t+1 forecast error, 1999 – 2008					
NNM	37,582	7.675	2.635	11.228	2.199
RW	37,582	1.016***	0.071**	2.995**	0.818***
HVZ	37,582	3.096***	1.683***	6.215***	1.796***
(d) t+1 forecast error, 2009 – 2017					
NNM	28,223	7.755	2.365	10.377	2.763
RW	28,223	1.374***	0.107***	15.230	1.014***
HVZ	28,223	2.423***	1.732***	3.094***	1.331***

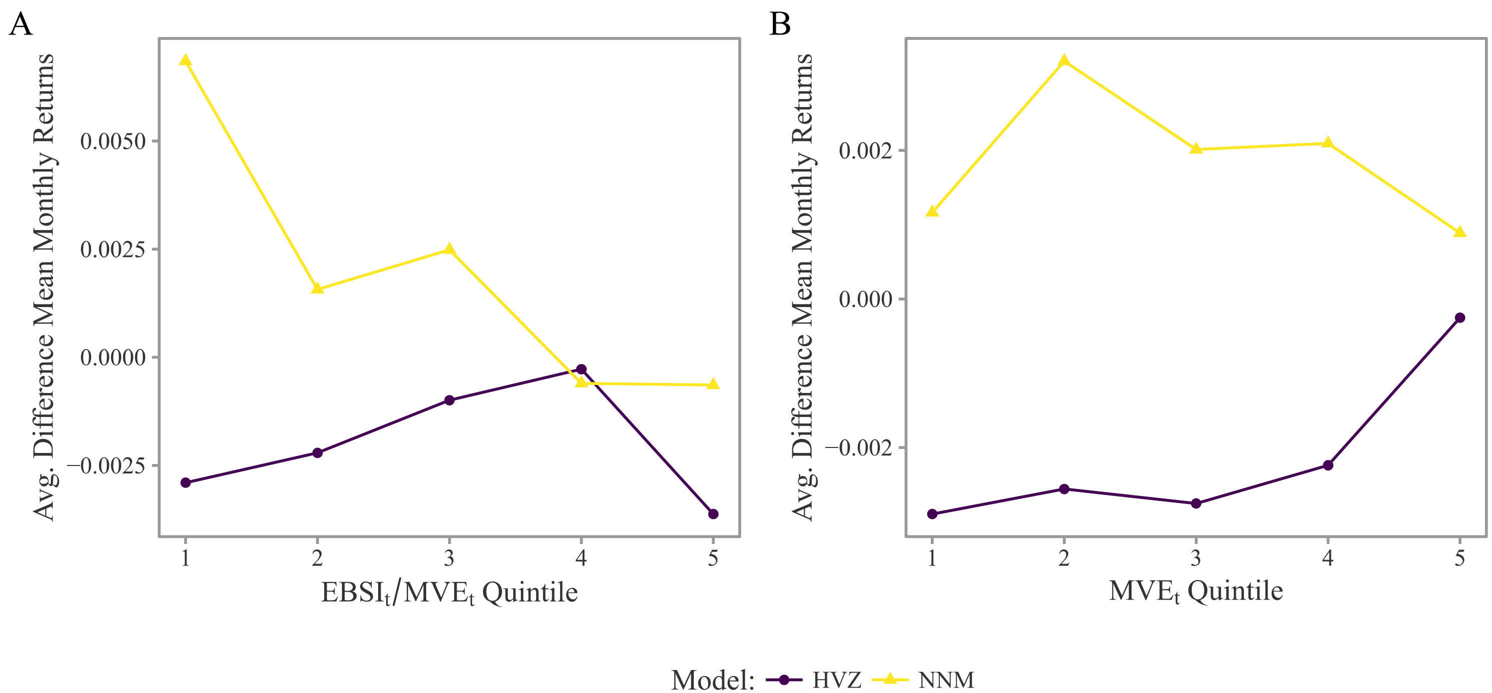
Cross-sectional differences in forecast accuracy



Industry differences in forecast accuracy



Differences in Future returns across partitions



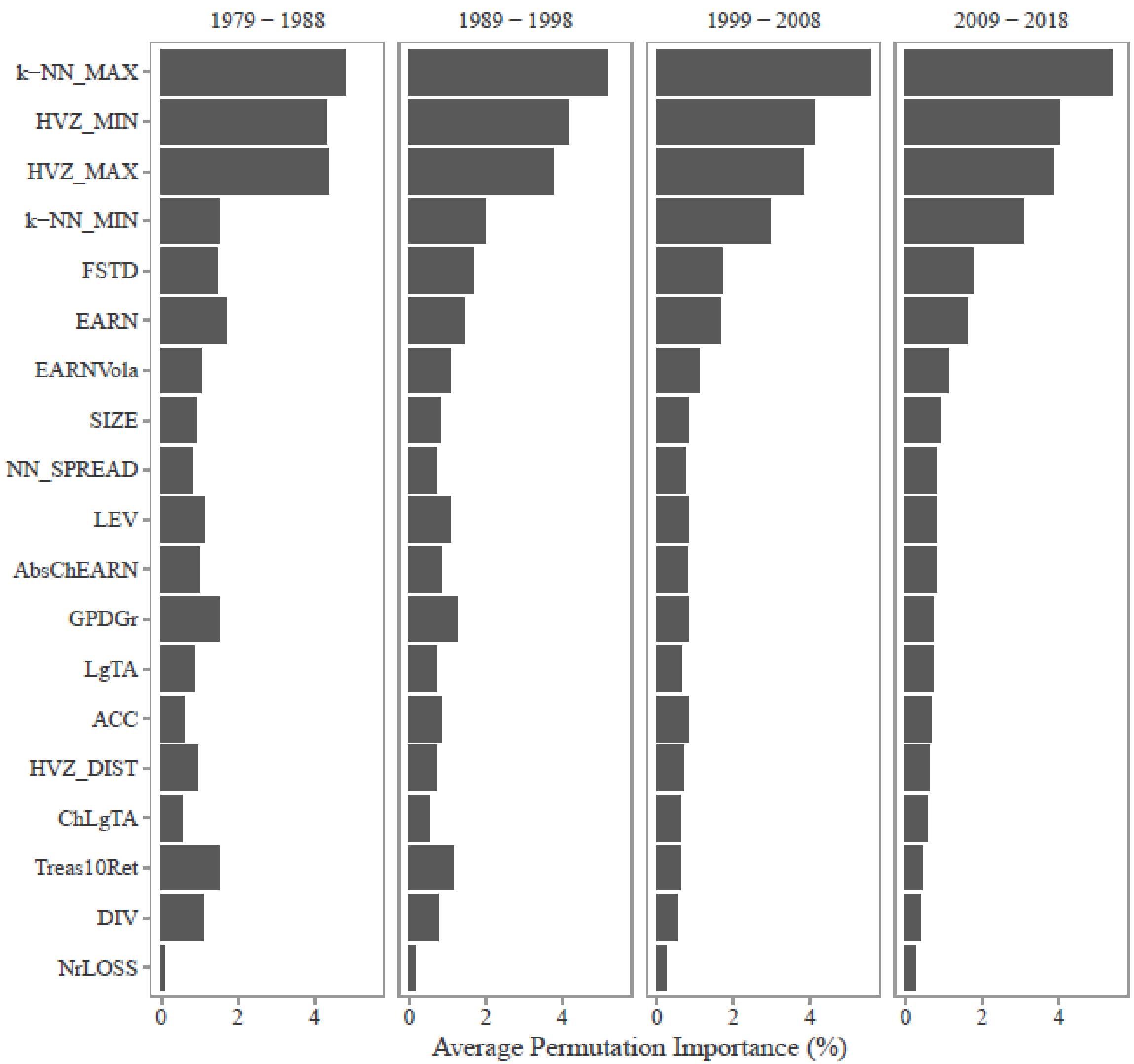
Can we predict the best model ex-ante?

Use a random forest classifier to predict ex-ante which of the three models is most accurate for a given firm-year:

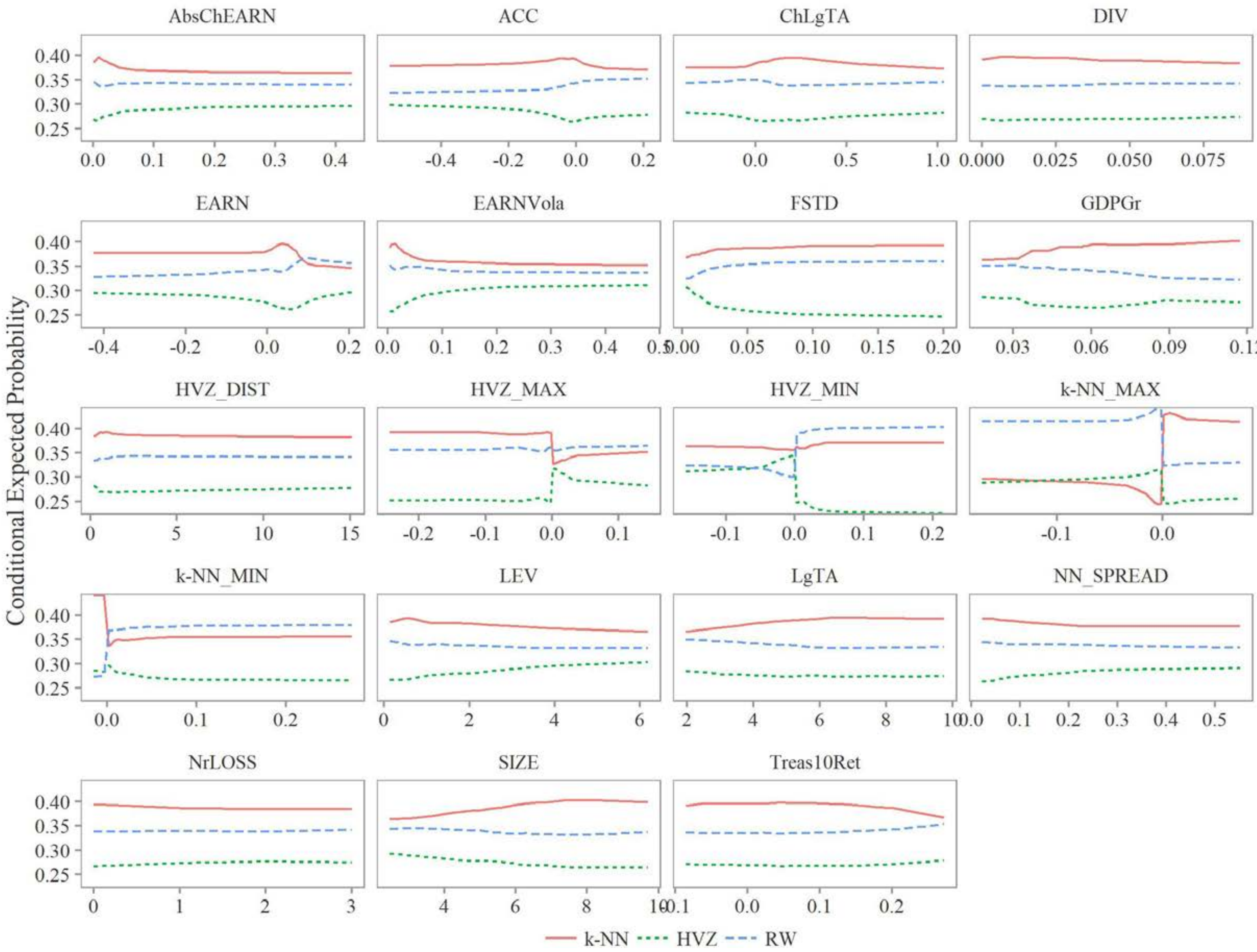
```
myTask <- makeClassifTask(data=x_joined, target="TopModel", check.data=FALSE)
myTask <- dropFeatures(myTask, c("gvkey", "calyear"))
myTask <- normalizeFeatures(myTask)

# Train learner
myLearner <- makeLearner("classif.ranger",
                         predict.type="prob",
                         importance="permutation",
                         num.trees=1000,
                         sample.fraction=1,
                         respect.unordered.factors="partition",
                         num.threads=5)
set.seed(123456, "L'Ecuyer-CMRG")
trainedLearner <- train(myLearner, myTask, subset=train_set)
```

What features predict the best model?



What features predict the best model?



Can we predict the best model ex-ante?

APPENDIX C: PREDICTING THE BEST MODEL

Table C.1: Predicting the best model to use by observation

Predicted	NNM	HVZ	RW	Total Predicted	Precision	Recall
NNM	32018	13069	14512	59599	0.537	0.678
HVZ	5305	9501	6402	21208	0.448	0.295
RW	9881	9629	20167	39677	0.508	0.491
Total Reference	47204	32199	41081	120484		
Accuracy	**0.512**					
No information rate	**0.392**					
Accuracy P-value	**0.00**					

Sort of?

Panel B: Overall forecast error comparison

Model	N	MAFE	MDAFE	MSE	WMSE	TMSE
NNM	120484	6.923	2.545	6.666	2.286	1.805
RW	120484	0.754***	0.085***	4.688	0.652***	0.444***
HVZ	120484	2.566***	1.461***	3.786***	1.540***	1.185***
RF	120484	0.122**	0.041**	0.458*	0.009	0.012

Table C.1 shows predicted outcomes as per the random forest and realized outcomes for each classified model. The rows of the matrix correspond to out-of-sample classifications as per the random forest. The columns correspond to the realized ("true") outcomes (classifications). Recall reflects the fraction of observations that belong to a class and are correctly classified. Precision reflects the fraction of predicted classifications that are correct.

Conclusion

- Default NNM model has the least data requirements but consistently better accuracy over all horizons
- Biggest gain in accuracy for dynamic/growing companies
- KNN approach puts earnings information into context better than other approaches
- Future research: explore these qualities further + examine properties of the spread in comparables

Thank you for your attention!