Predicting NCAAB match outcomes

(a numbers talk)

Zifan Shi, Sruthi Moorthy, A. Zimmermann

Who wins a b-ball game?

The team that scores more points!

How do you score points?

You put the ball into the basket!

How do you score points?

You put the ball into the basket!

- Shots: 2 pointers, 3 pointers, free throws

How do you score points?

You put the ball into the basket!

- Shots: 2 pointers, 3 pointers, free throws
- Get/keep the ball: steals, rebounds

How do you score points?

You put the ball into the basket!

- Shots: 2 pointers, 3 pointers, free throws
- Get/keep the ball: steals, rebounds
- Don’t lose the ball: turnovers, fouls

Normalization (1)

10/50 possible rebounds - not that great

10/30 possible rebounds - much better

Oliver 2004

Four Factors

1. Effective field goal percentage
2. Offensive rebound rate
3. Turnover rate
4. Free throw rate

Normalization (2)

Take pace into account: slower games ⇒ **fewer** points

Efficiencies:

- Count possessions, divide by possessions, multiply by 100

Normalization (3)

Take opponent into account: many points against weak defensive team ⇒ **not** impressive

Adjusted efficiencies:

- Divide by opponent's counter stat, multiply by overall average

Pomeroy

Non-ML predictions

Average over season so far

Pythagorean expectation:

$$WinProbability = \frac{((Adjusted)OE_{avg})^y}{((Adjusted)OE_{avg})^y + ((Adjusted)DE_{avg})^y}$$

Naive Zimmermann assumption

Naive Zimmermann assumption

Using Machine Learning instead of statistical methods will lead to better results.

Setting

NCAAB match data

- Location, normalized attributes, win/loss

Season	2009	2010	2011	2012	2013
Train	5265	10601	15990	21373	26772
Test	5336	5389	5383	5399	5464

Prequential error

Season	J48	RF	NB	MLP
2009	68.4%	68.8%	71.1%	70.8%
2010	68.9%	69.4%	71.7%	72.5%
2011	69.1%	67.8%	70.3%	71.6%
2012	70.4%	71.4%	72.8%	74.5%
2013	68.9%	68.8%	71.9%	72.2%

Overfitting: default parameters at fault

Season	J48	RF	NB	MLP
2009	68.4%	68.8%	71.1%	70.8%
2010	68.9%	69.4%	71.7%	72.5%
2011	69.1%	67.8%	70.3%	71.6%
2012	70.4%	71.4%	72.8%	74.5%
2013	68.9%	68.8%	71.9%	72.2%

Season	J48	RF	NB	MLP
2009	68.4%	68.8%	71.1%	70.8%
2010	68.9%	69.4%	71.7%	72.5%
2011	69.1%	67.8%	70.3%	71.6%
2012	70.4%	71.4%	72.8%	74.5%
2013	68.9%	68.8%	71.9%	72.2%

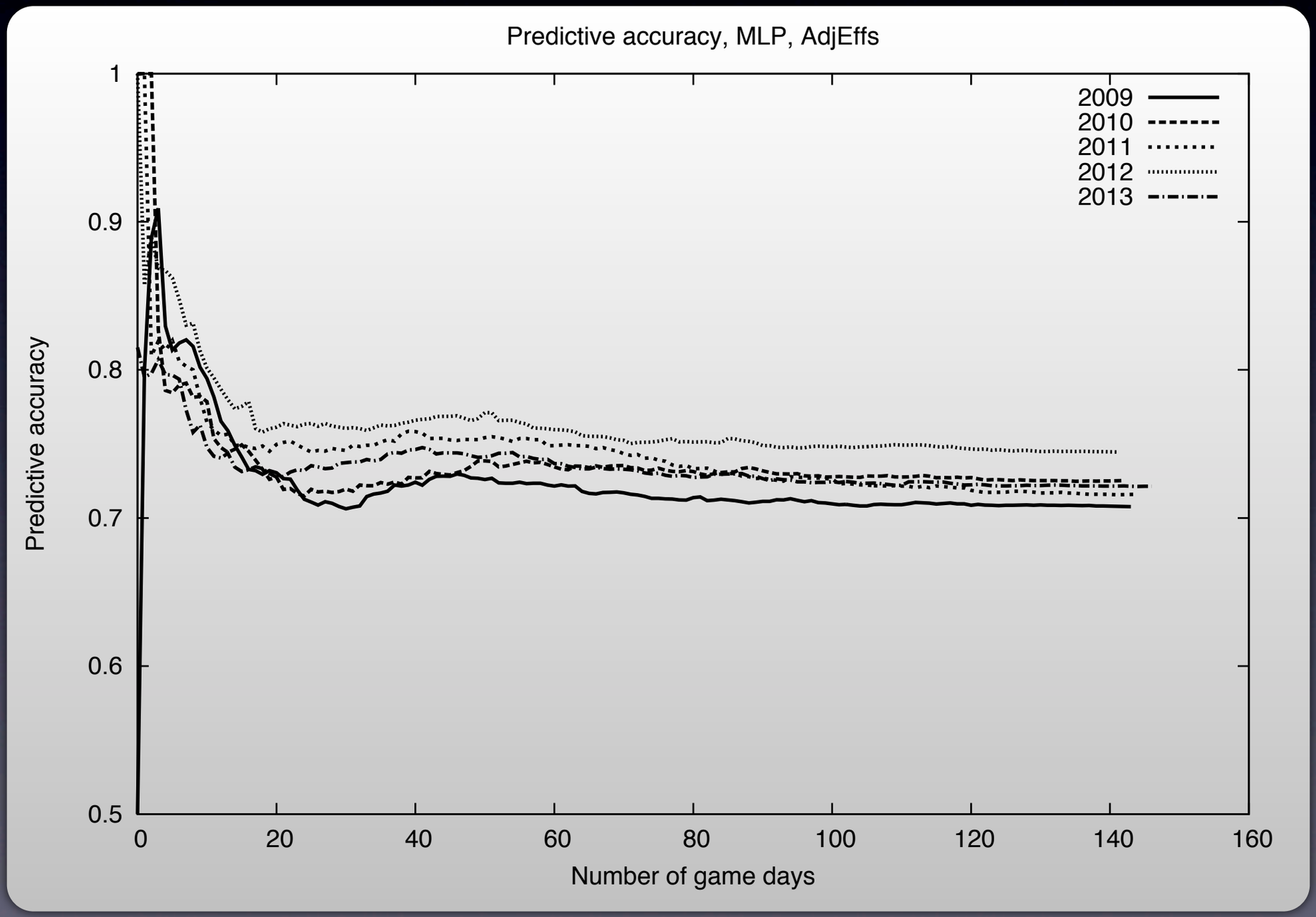
More training data not better

Naive assumption doesn't hold!

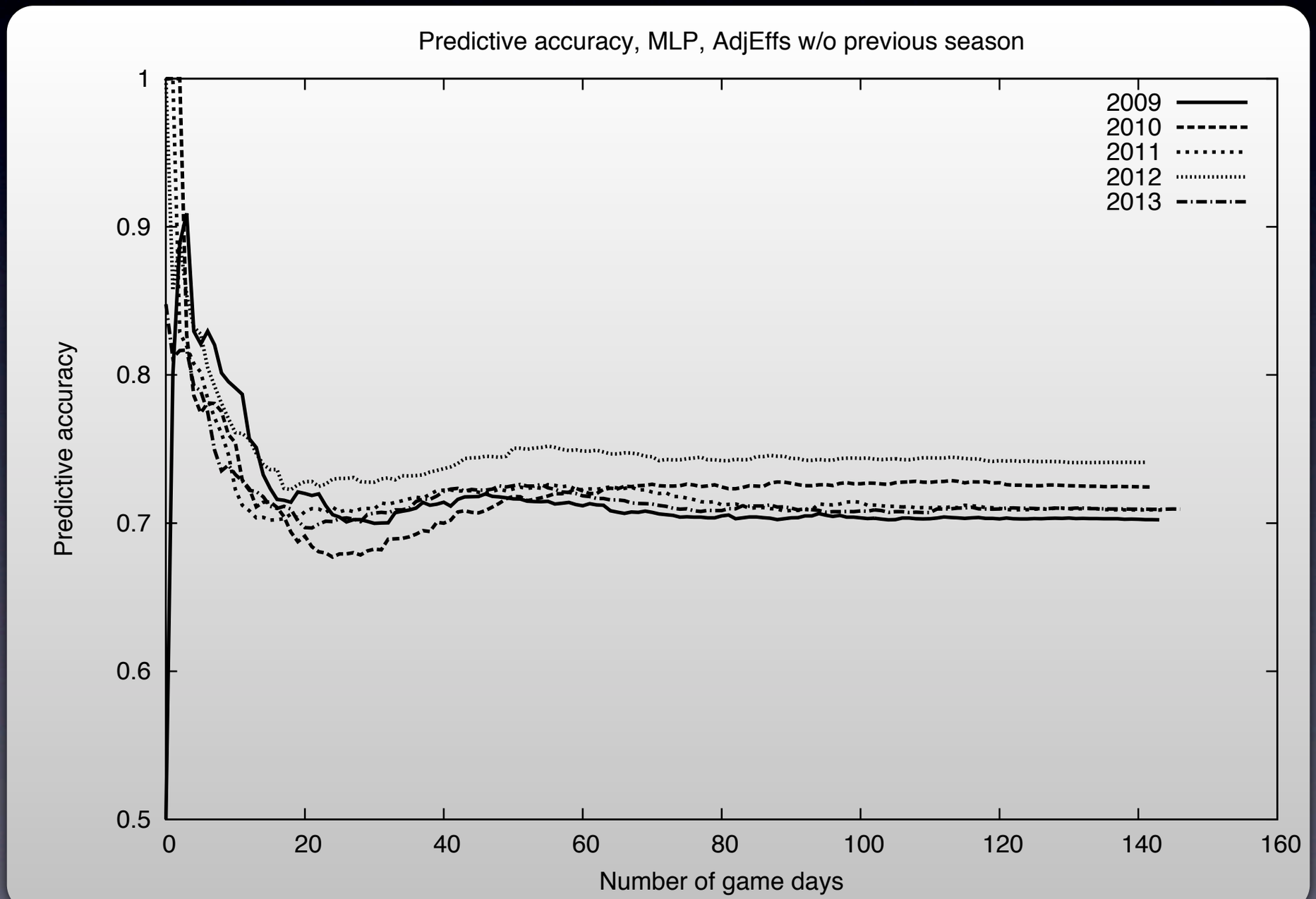
Season	J48	RF	NB	MLP
2009	68.4%	68.8%	71.1%	70.8%
2010	68.9%	69.4%	71.7%	72.5%
2011	69.1%	67.8%	70.3%	71.6%
2012	70.4%	71.4%	72.8%	74.5%
2013	68.9%	68.8%	71.9%	72.2%

Lesson 1: it's in the attributes

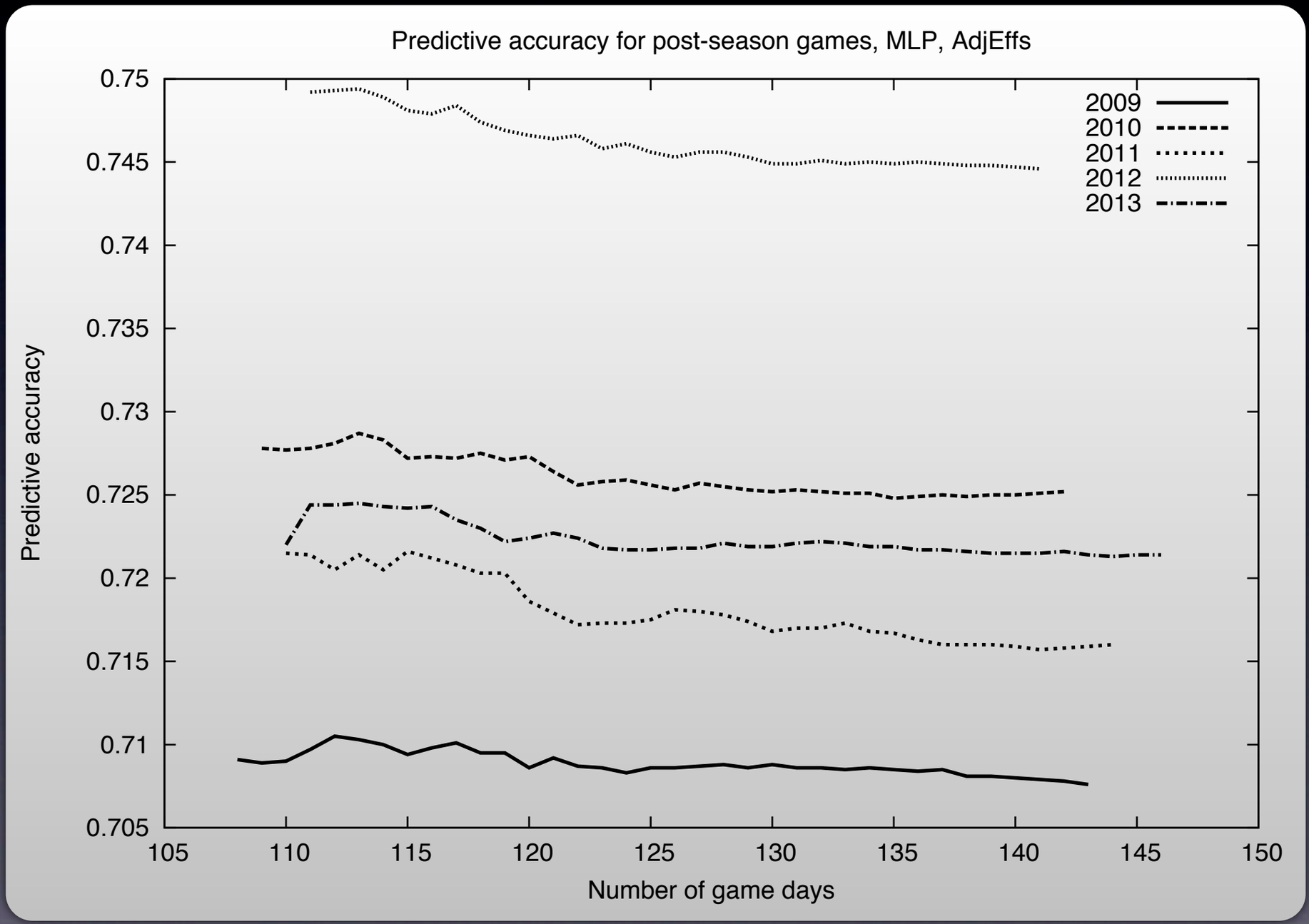
Adjusted Efficiencies, MLP



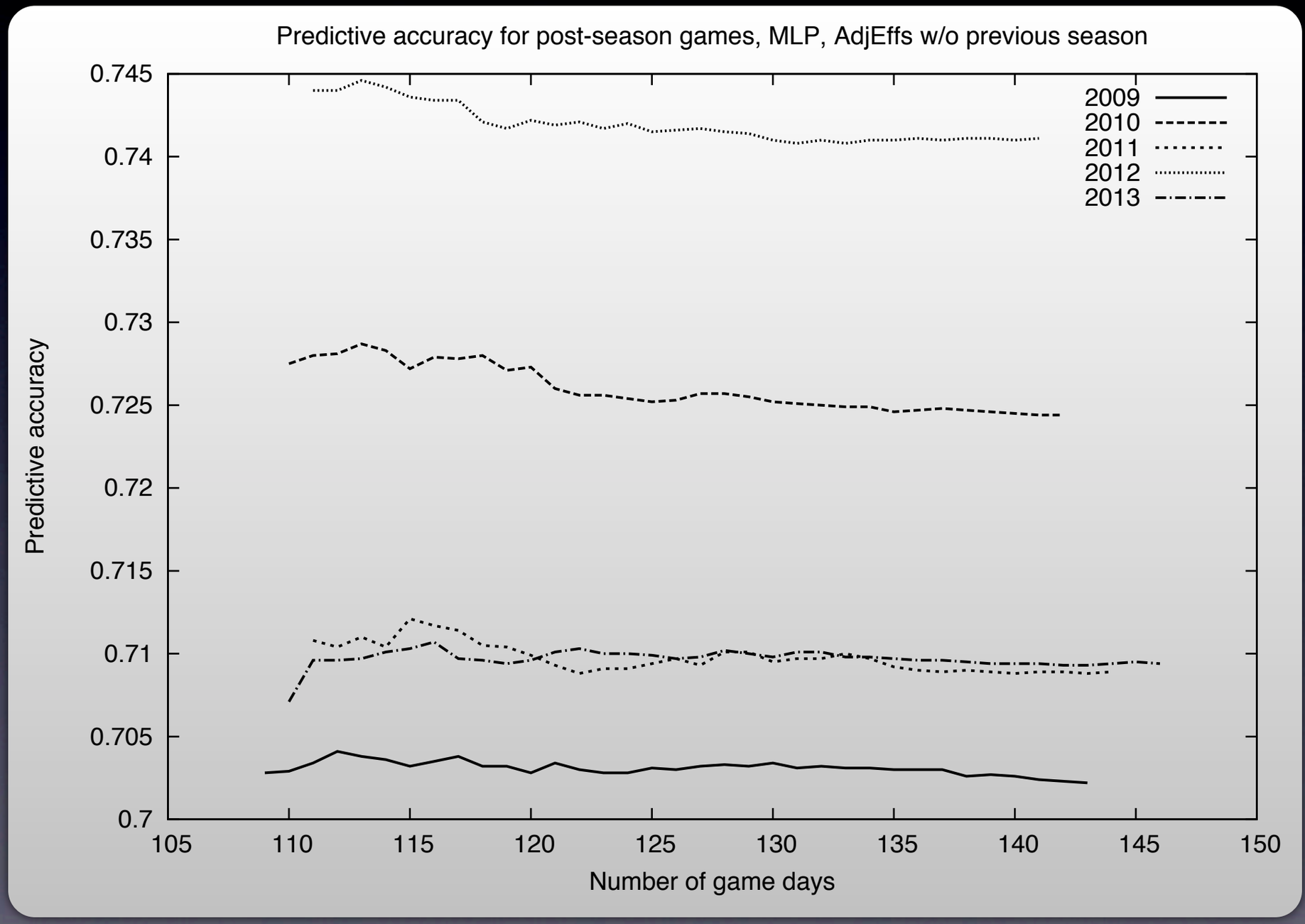
Adjusted Efficiencies, MLP



AdjEff, MLP, Post-season

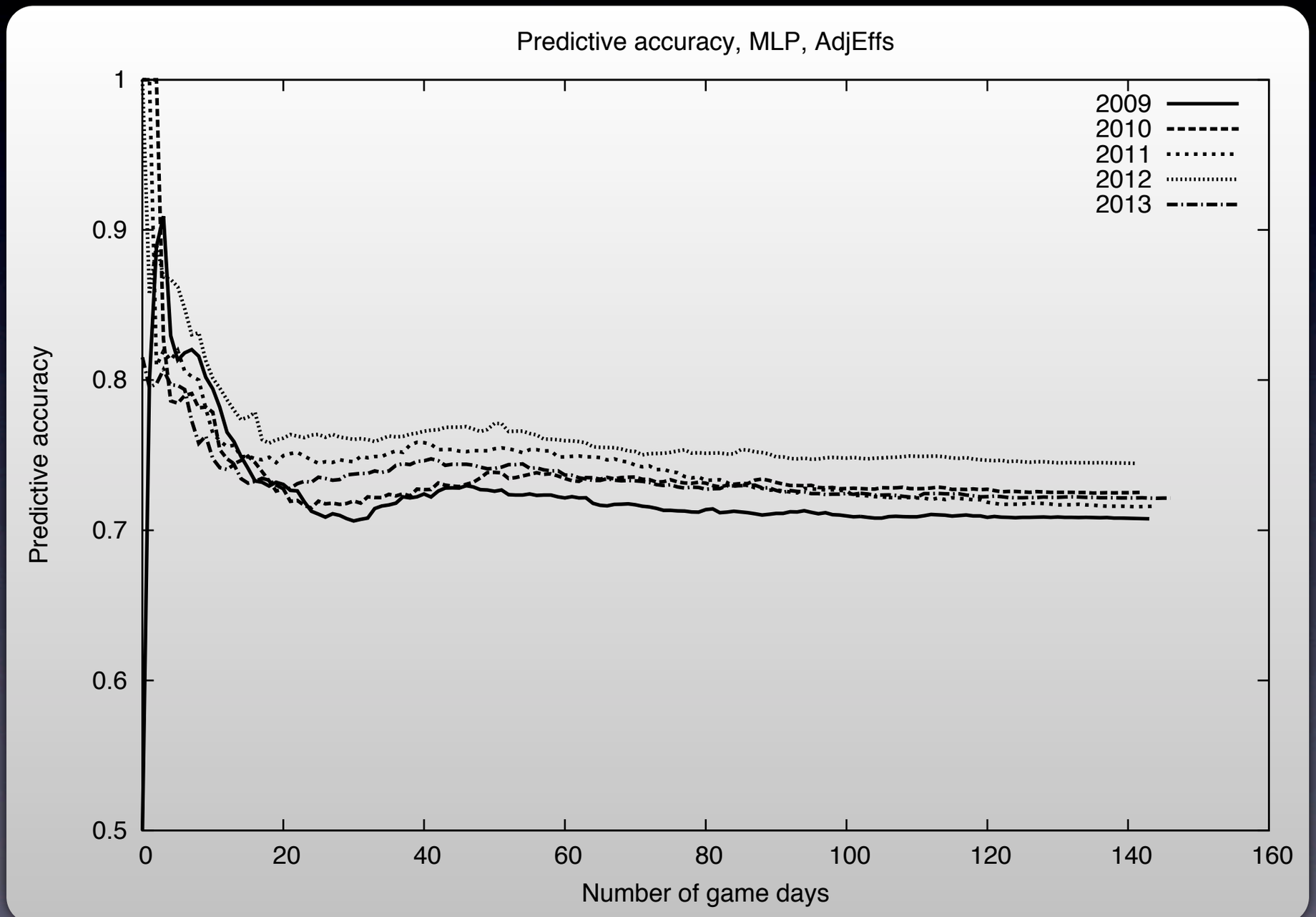


AdjEff, MLP, Post-season



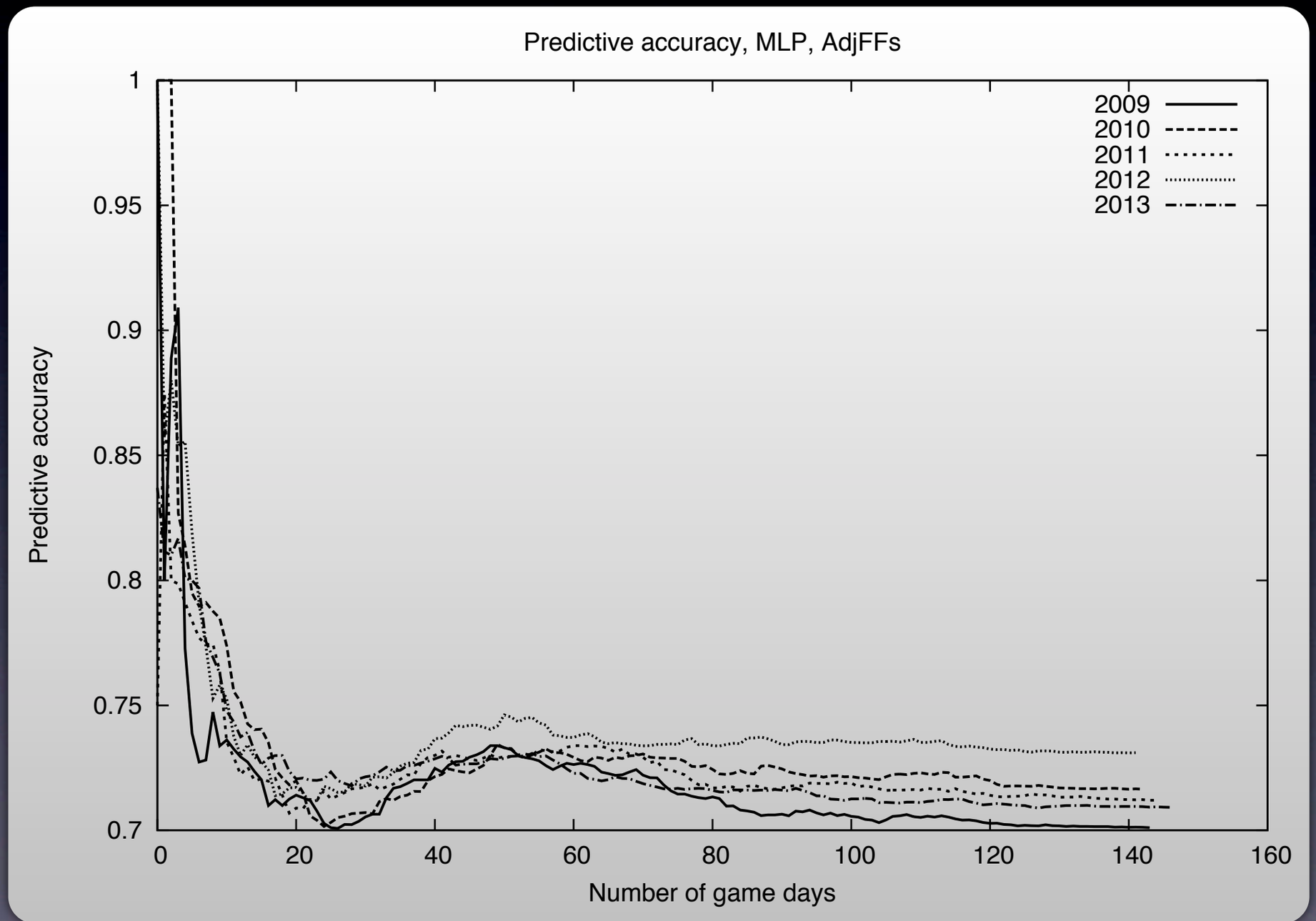
Contribution :)

AdjFF, MLP



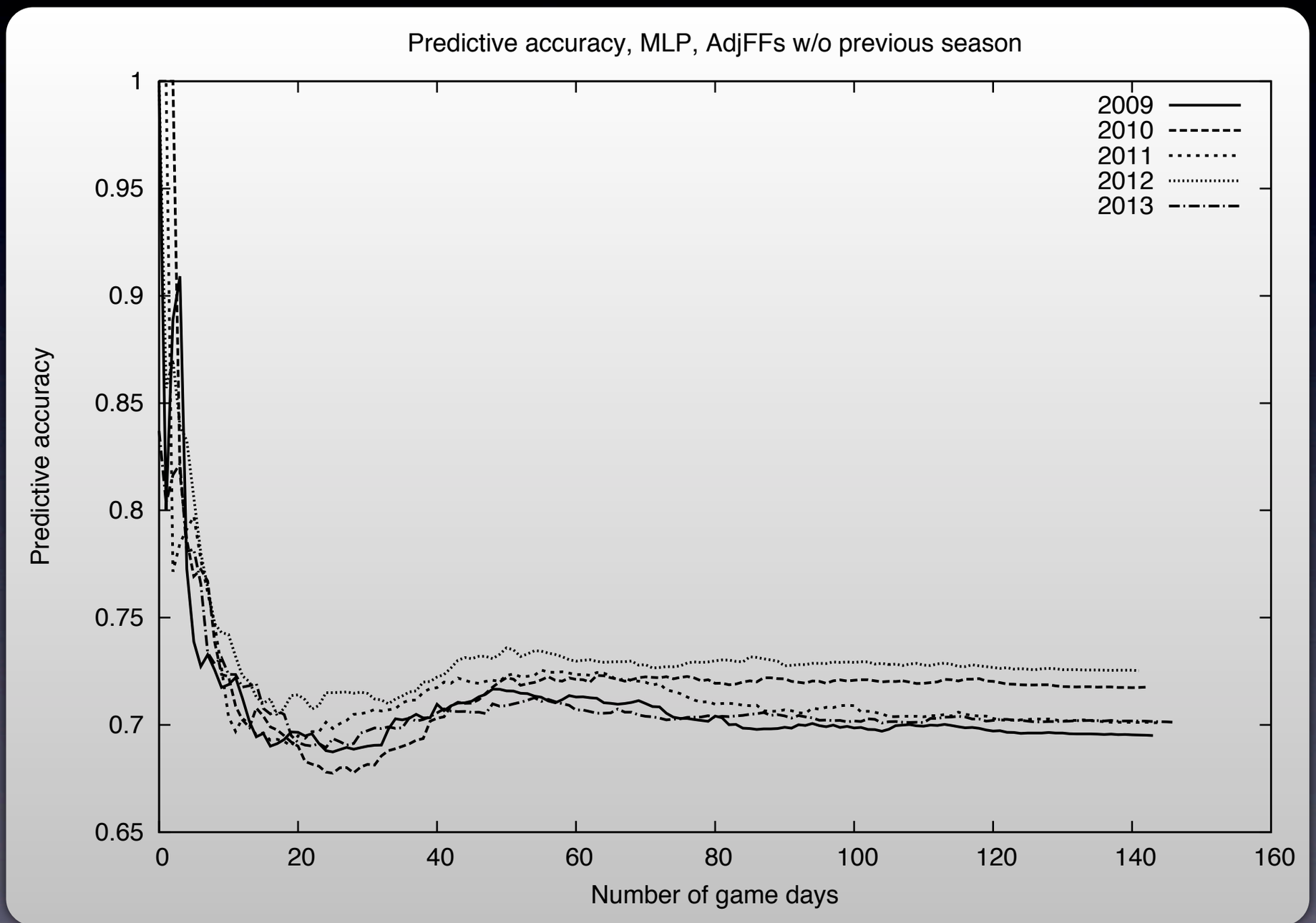
Contribution :)

AdjFF, MLP

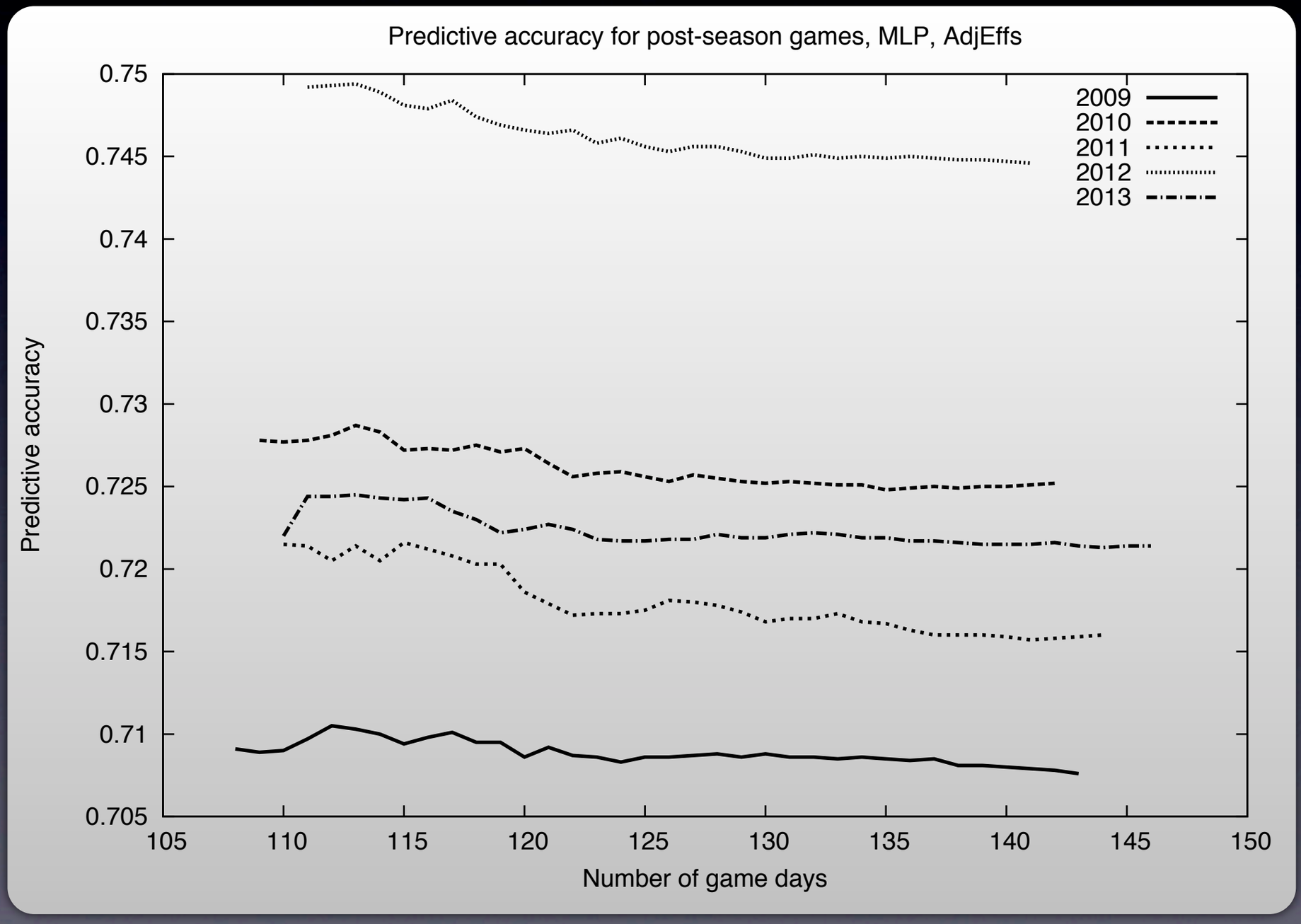


Contribution :)

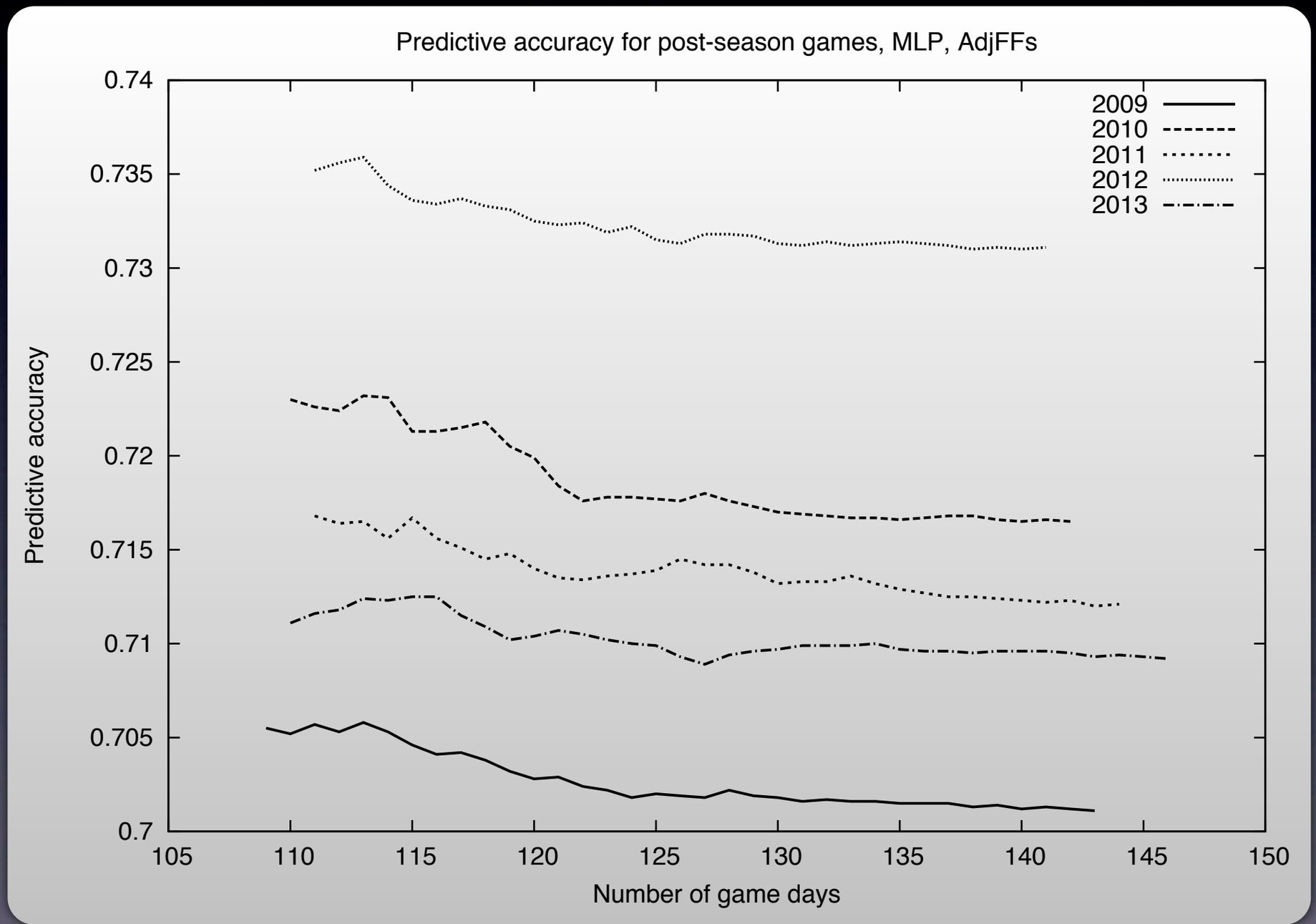
AdjFF, MLP



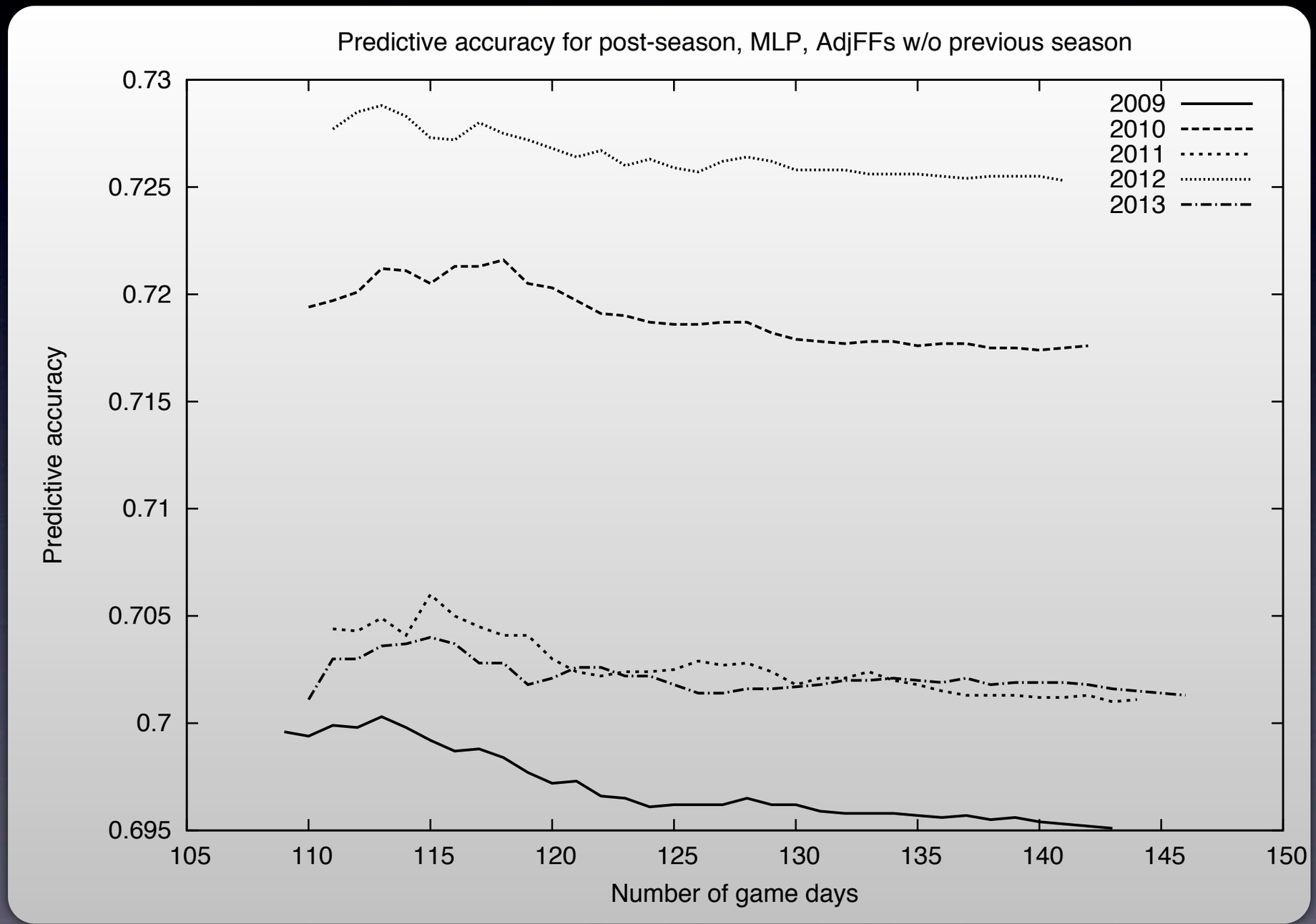
AdjFF, MLP, Post-season



AdjFF, MLP, Post-season

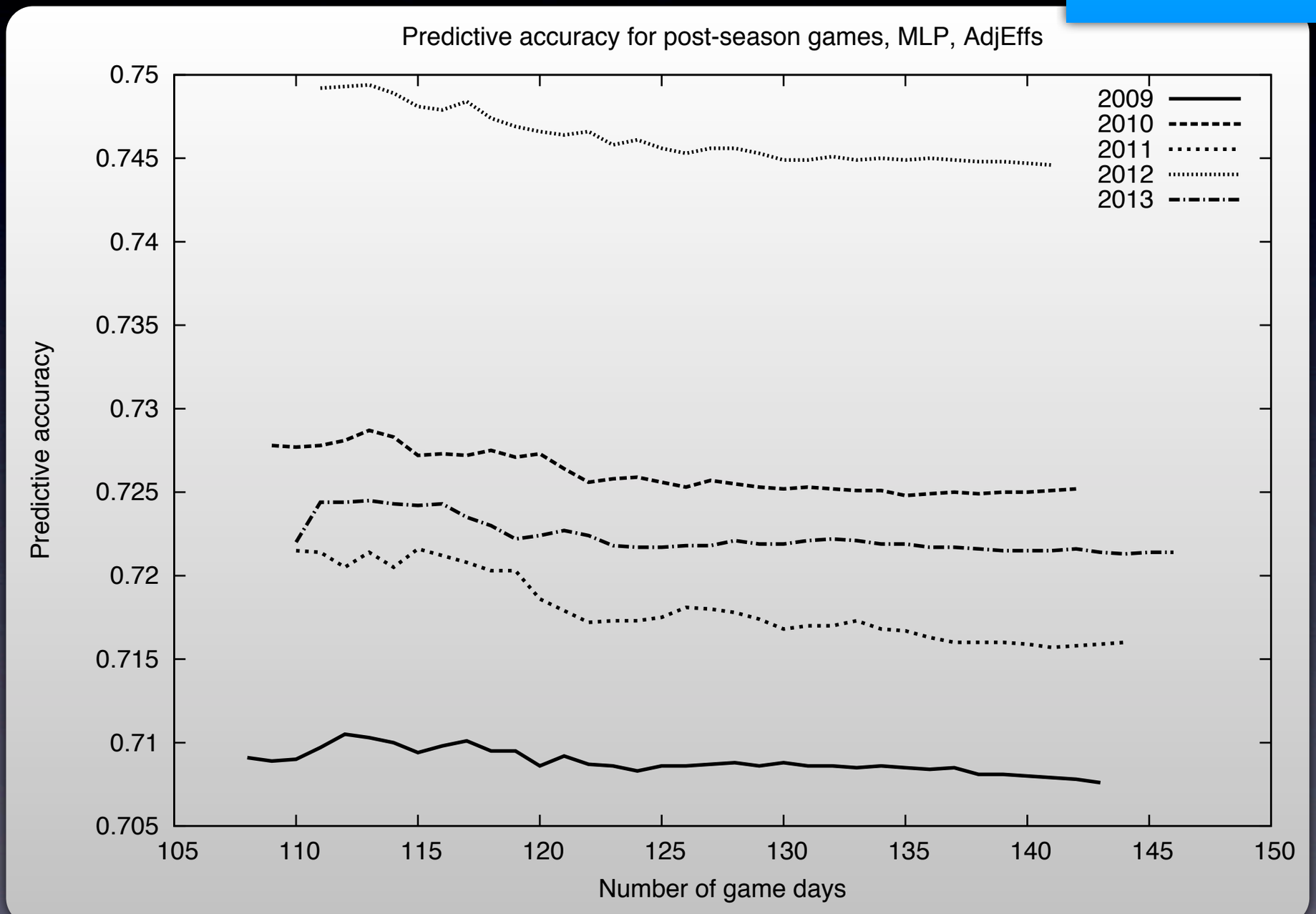


AdjFF, MLP, Post-season



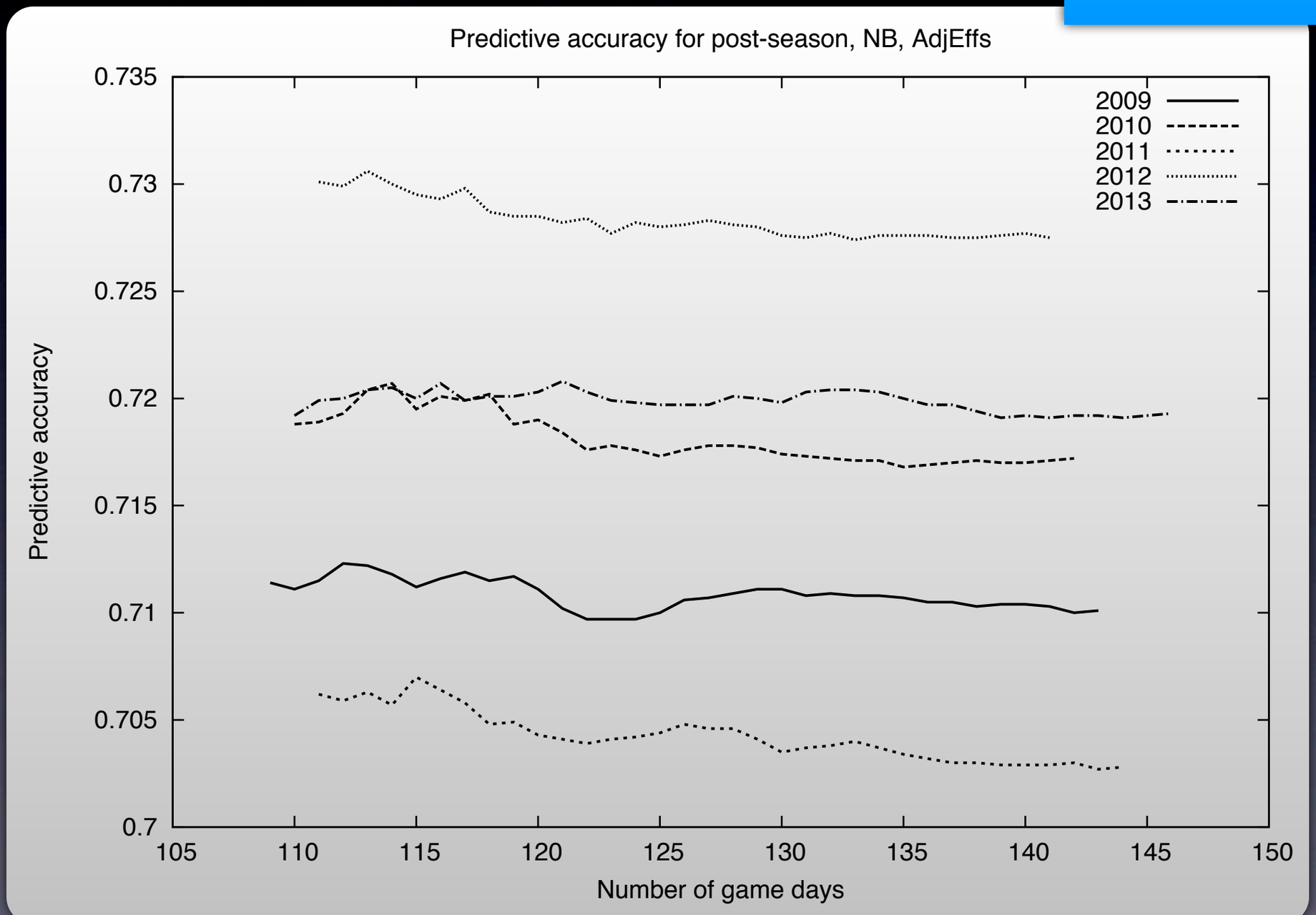
Naive Bayes

Maybe a bit on the technique .



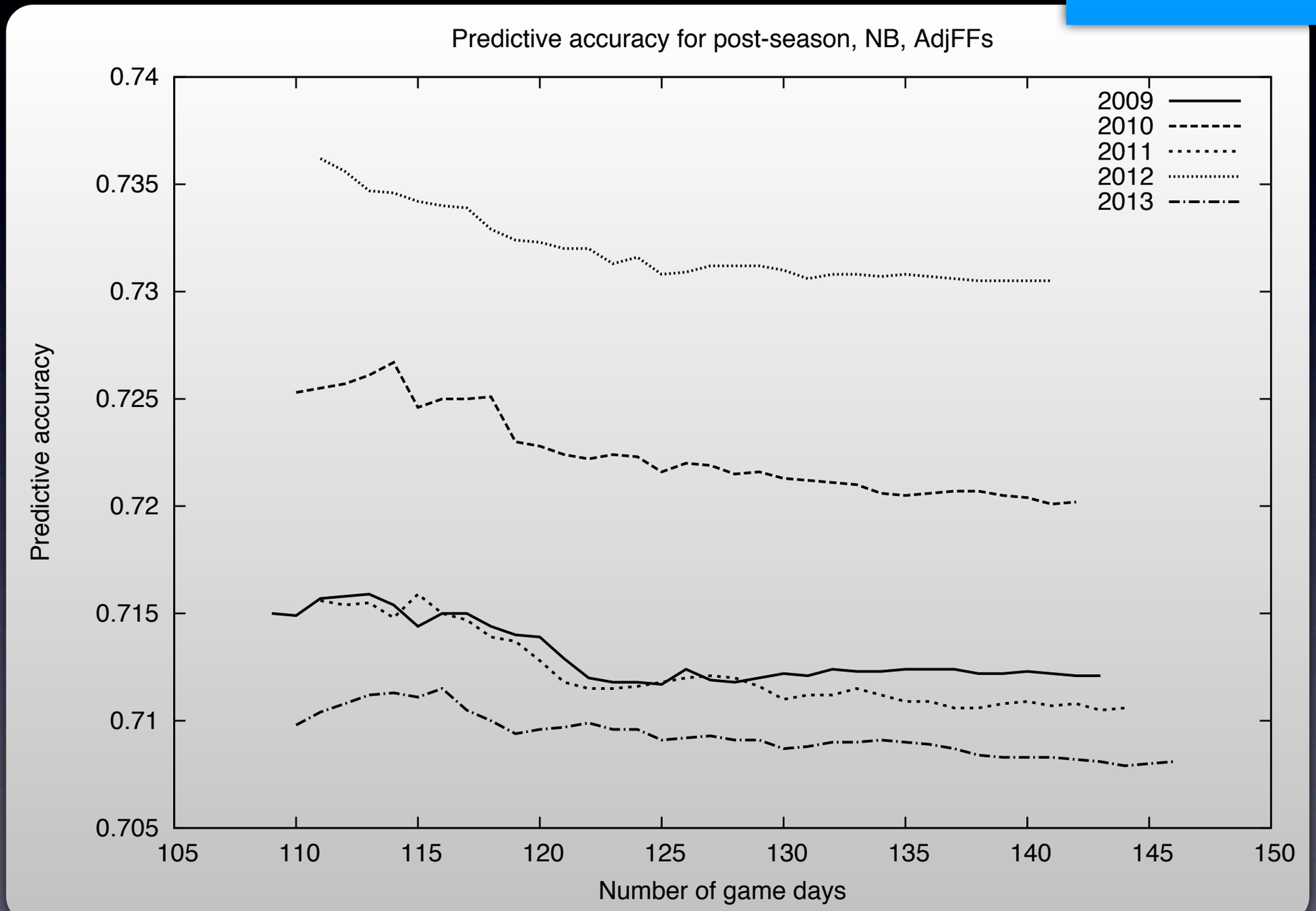
Naive Bayes

Maybe a bit on the technique .



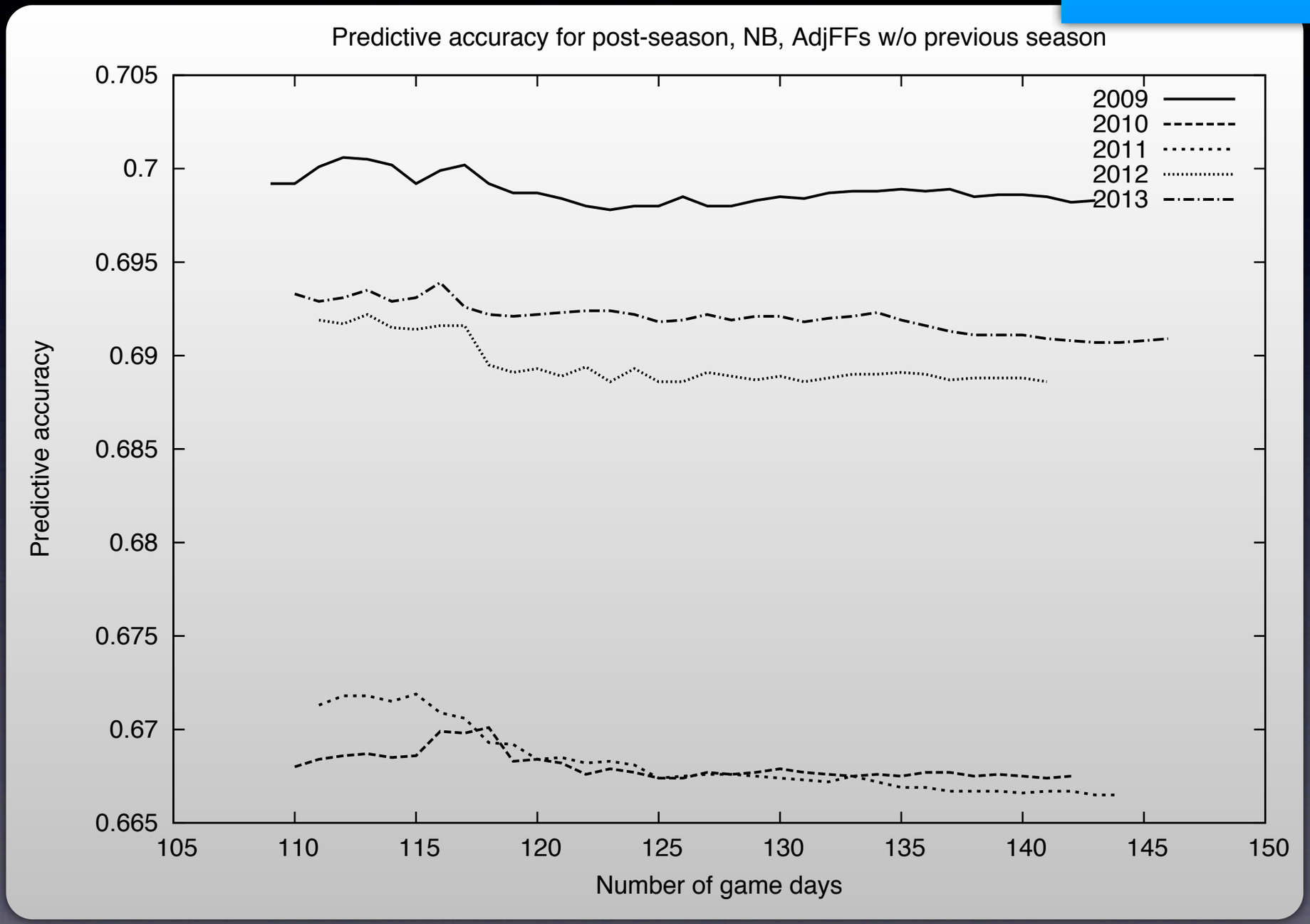
Naive Bayes

Maybe a bit on the technique .



Naive Bayes

Maybe a bit on the technique .



Lesson 2: there might be a class ceiling

- ~ 75% rarely exceeded
- Holds for other sports as well (NHL, Soccer, NFL)

Next steps

Better attributes

- Team stability/experience

Focus on mis-classifications

Sequential classification

- Concept drift?

Next steps

Better attributes

- Team stability/experience

Focus on mis-classifications

Sequential classification

- Concept drift?

Season	MLP
2009	70.8%
2010	72.5%
2011	71.6%
2012	74.5%
2013	72.2%

More lessons?