Exploring Chance in NCAA basketball

Albrecht Zimmermann
CoDaG, GREYC, Université Caen

MLSA @ ECML/PKDD '15 – 11/09/2015

UNICAEN
UNIVERSITÉ
CAEN
NORMANDIE

Problem

NCAAB match outcome predictions at most 76 %

Similar phenomena in other sports

Is there a limiting factor ?

- Intuition : there is – chance in match outcomes

Modelling chance (not my idea)

Assumption : win counts/probabilities follow Poisson distribution

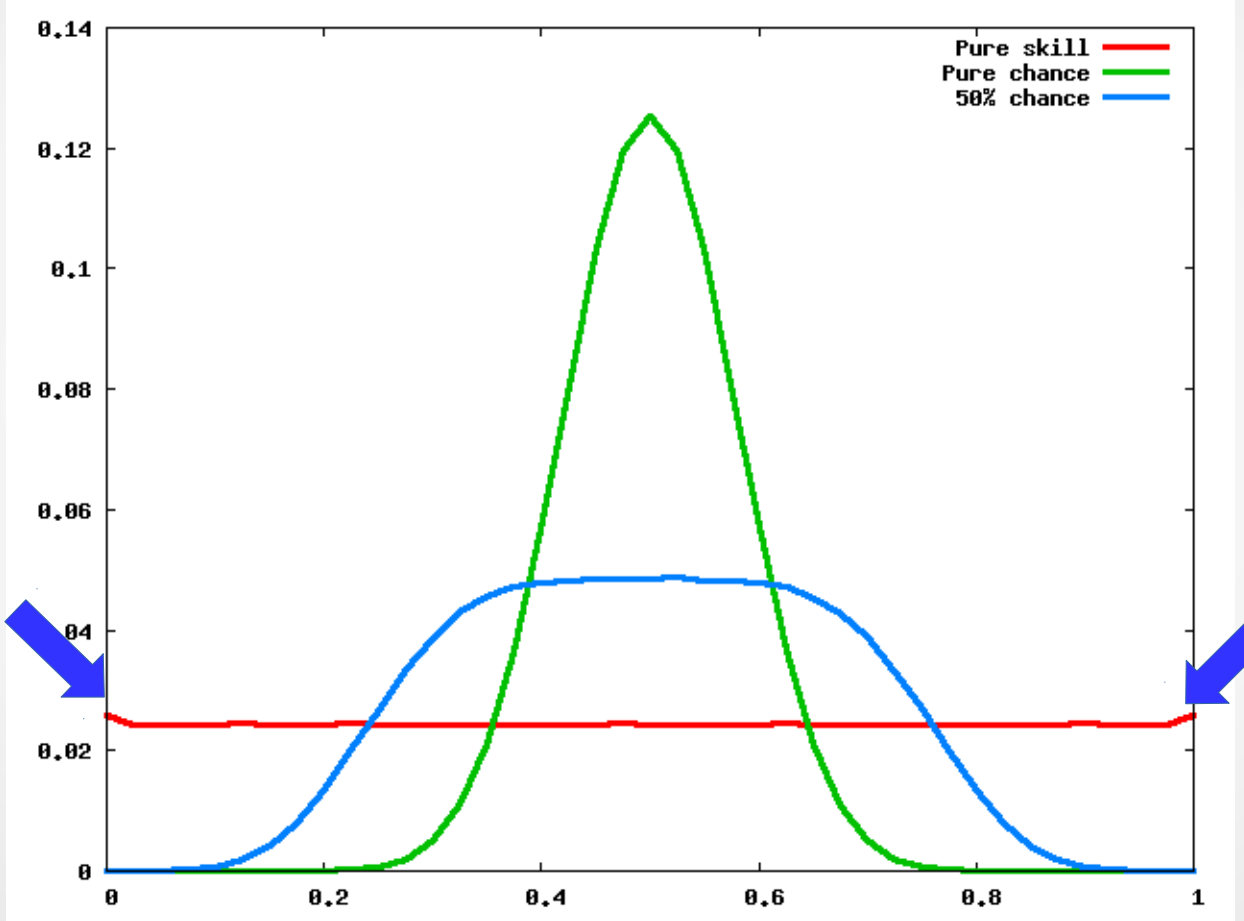
Approximation by MC simulation

- Pick team strengths randomly
- Set chance parameter c
- Match teams randomly
- In 1-c cases stronger team wins, in c coin flip

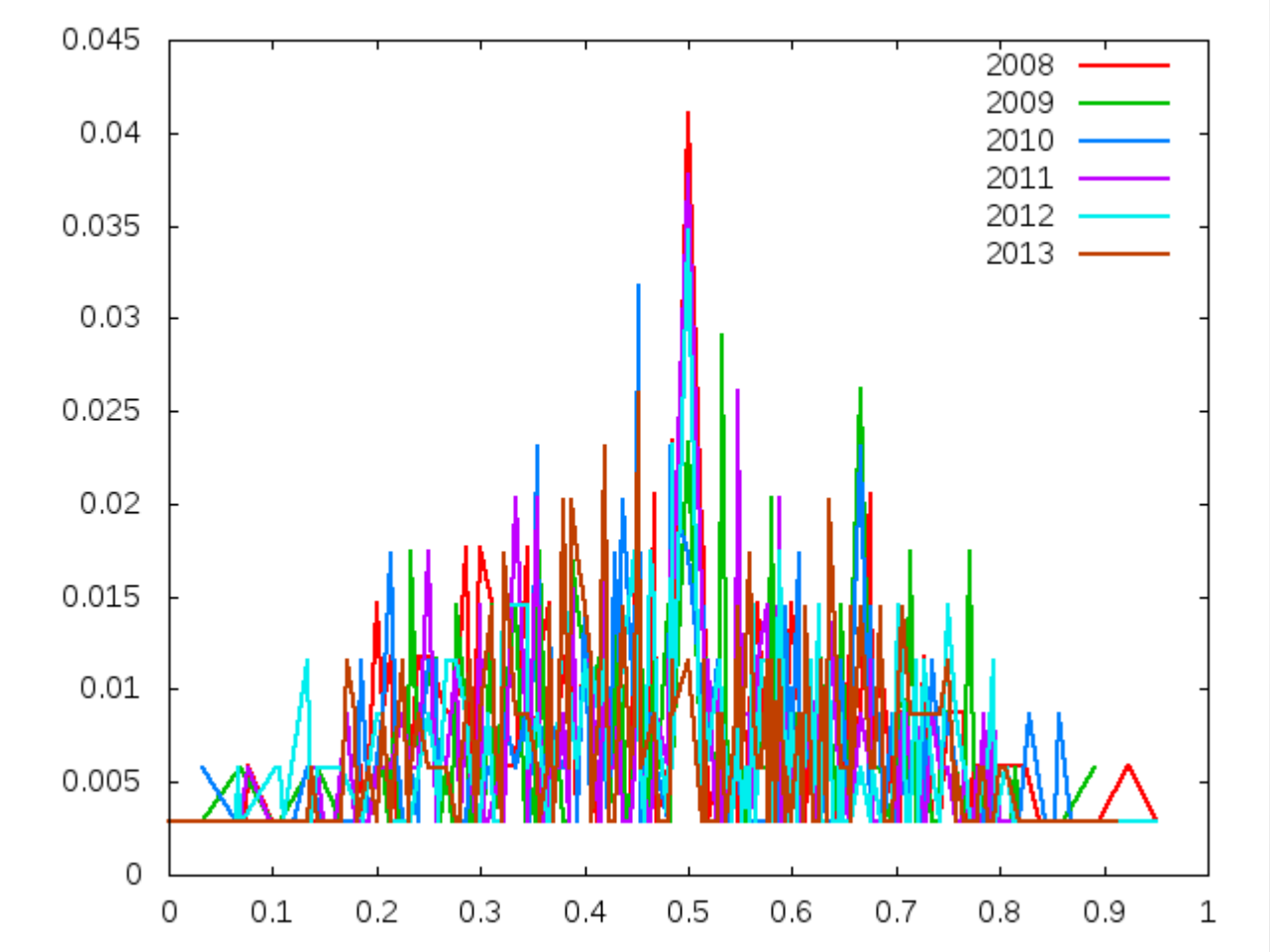
Repeat 10k times

Pick c for which distribution is « closest »

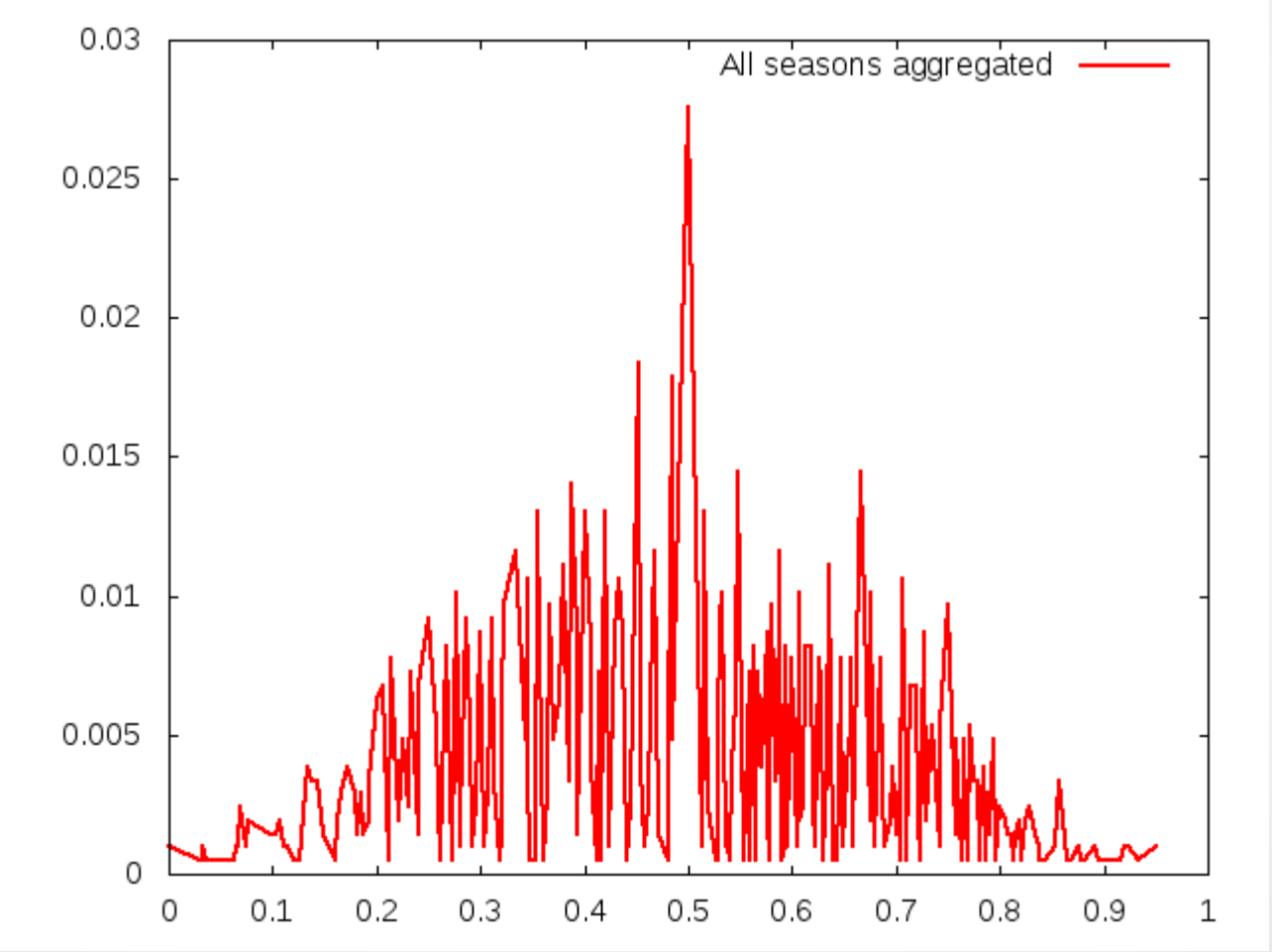
MC results


Pure skill
Pure chance
50% chance
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
0
0.2
0.4
0.6
0.8
1

Observed results, NCAAB



Observed results



Issues with unbiased MC

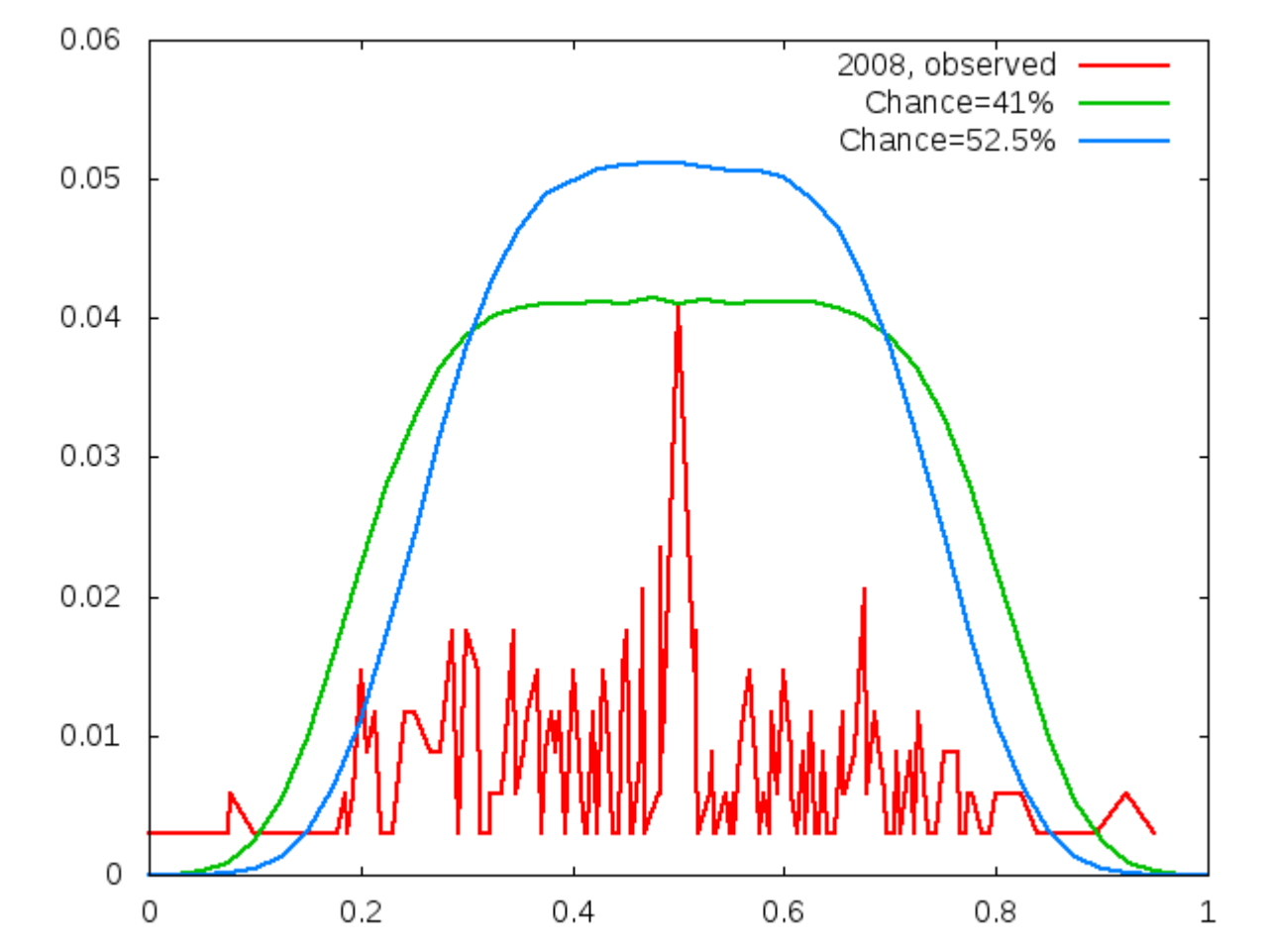
MC sim

Effects

MC lon

season

ndividual



Biased chance modelling (my idea)

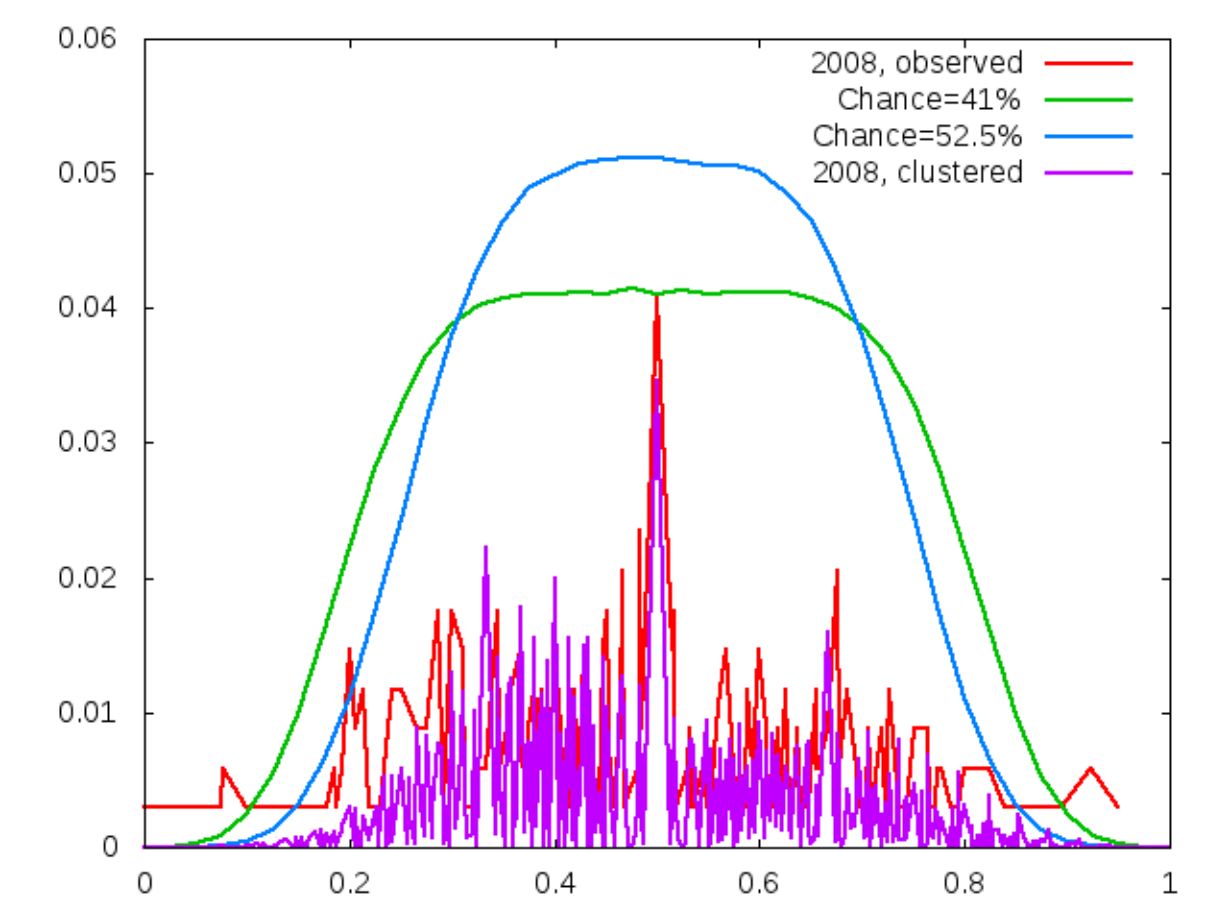
Assumption 1 : not all teams play a balanced schedule

Assumption 2 : chance has larger effect in match-ups of similar teams

Assumption 3 : « similar » can be found by clustering team statistical profiles

- No assignment team strengths
- c not chosen but read from season's cluster match-ups
- # clusters could be chosen ; EM can do on its own

Biased chance simulation



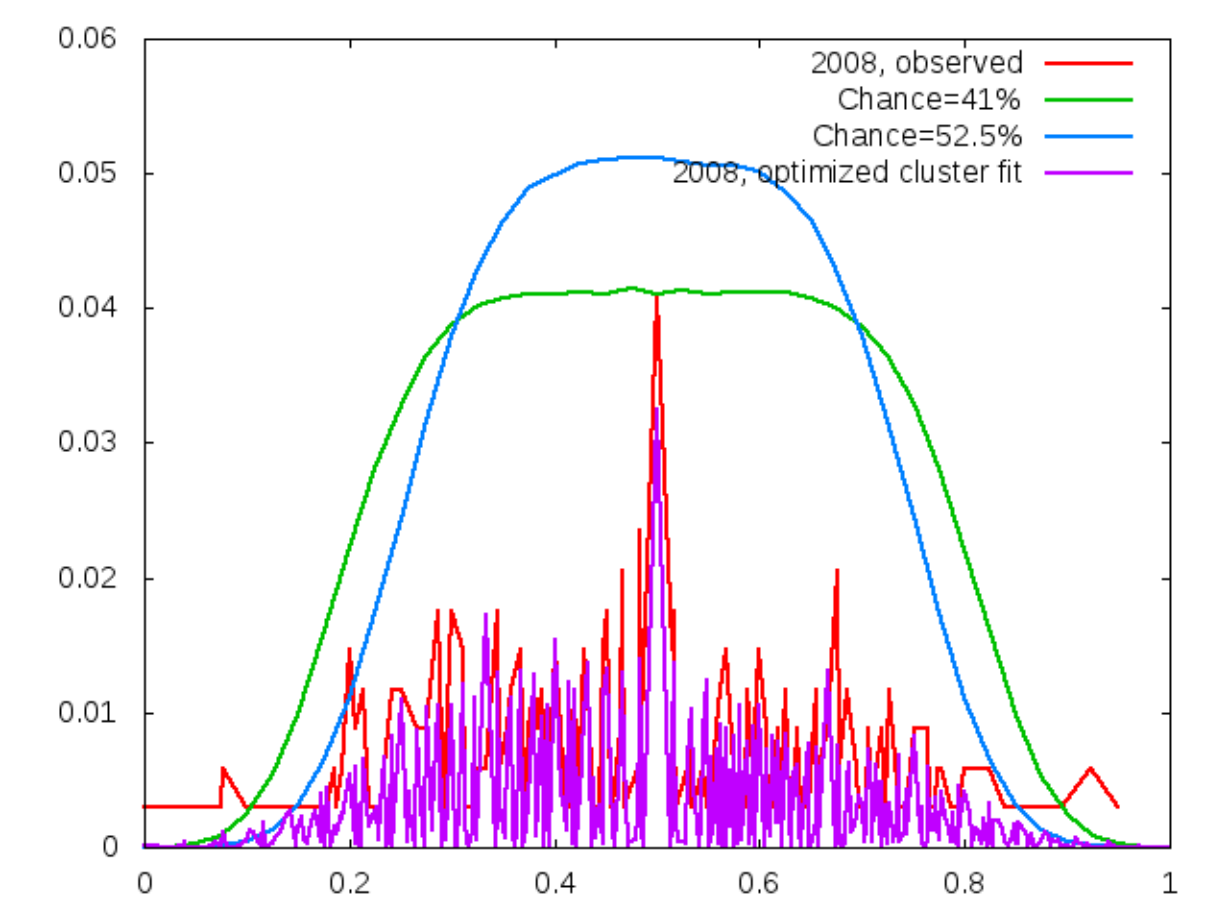
Chance estimates

Season	2008	2009	2010	2011	2012	2013
Chance	0.5736	0.5341	0.5066	0.5343	0.5486	0.5322
Predictive Limit	0.7132	0.7329	0.7467	0.7329	0.7257	0.7339
KenPom	0.7105	0.7112	0.7244	0.7148	0.7307	0.7035

EM-found clusters might not be appropriate

→ Find clustering leading to « closest » distribution

Revised biased chance simulation



Revised chance estimates

Season	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**
Chance	0.4779	0.5341	0.4704	0.5343	0.4853	0.5311
Predictive Limit	0.7610	0.7329	0.7648	0.7329	0.7573	0.7345
KenPom	0.7105	0.7112	0.7244	0.7148	**0.7307**	0.7035

What to do with this information?

1. Exploration of other sports (NBA done, NFL next)
   - And other league set-ups (remember Jan's talk)
2. Performance assessment
3. Maybe direct classifier (non-deterministic)
4. Data generation (filling in the gaps)

albrecht.zimmermann@unicaen.fr