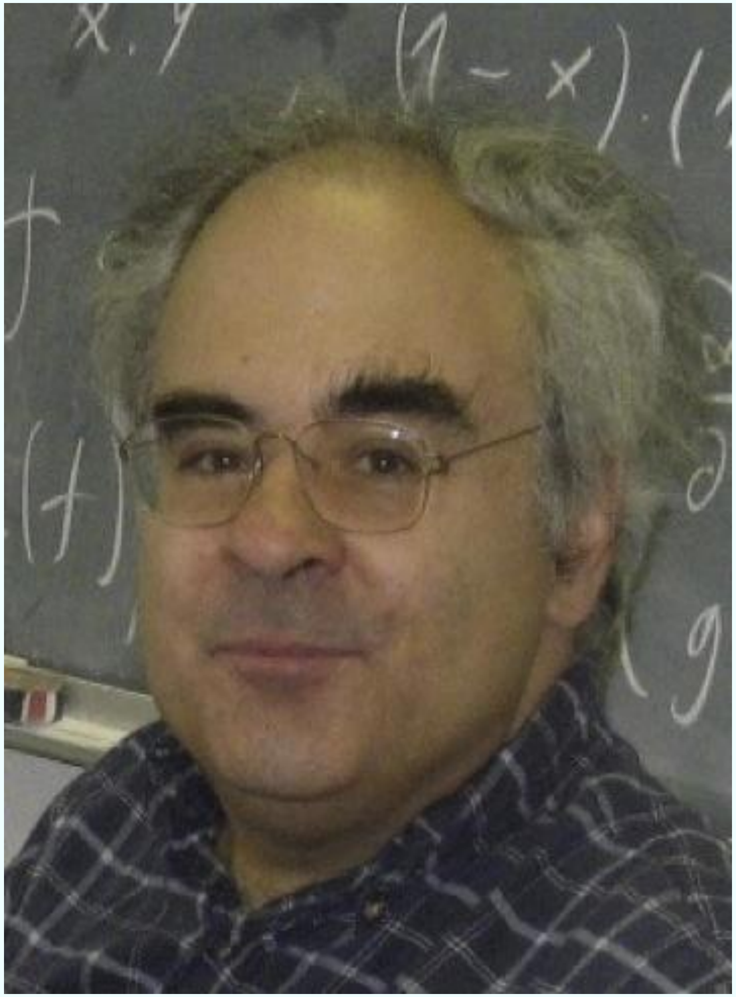
Understanding distributions of chess performances

Understanding Distributions of Performance

Dr. Kenneth W. Regan, Univ. at Buffalo
Bartlomiej Macieja, Warsaw, Poland
Guy Haworth, Univ. of Reading (UK)

ACG13, Tilburg, The Netherlands, Nov. 2011



Kenneth Regan

Bartlomiej Macieja

Back story

- **A sequence of papers on ‘Assessing Chess Players’**
  - **Reference Fallible Endgame Players (2002, 2003)**
  - **(Deeper) Model Endgame Analysis (2003, 2005)**
  - **Reference Fallible Players (2007)**
  - **Skill Rating by Bayesian Inference (2009)** ... IEEE CIDM ‘09
  - **Performance and Prediction, (2009)** ... ACG12, Pamplona
  - **Intrinsic Chess Ratings (2011)** ... AAAI-11, San Francisco
- **Topics**
  - **The creation of a Skilloscope to rank players**
  - **Comparison of and correlation with ELO scales**
  - **Detection of plagiarism ... and ELO Scale instability**

The focus today

- the question of *ELO Inflation*
- common views about the FIDE ELO scale
  - View 1: *ELO 2700* means *lower quality play* today
  - View 2: *ELO 2700* should mean ‘best few’ players
  - it is impossible for ELO to conform to both views over time

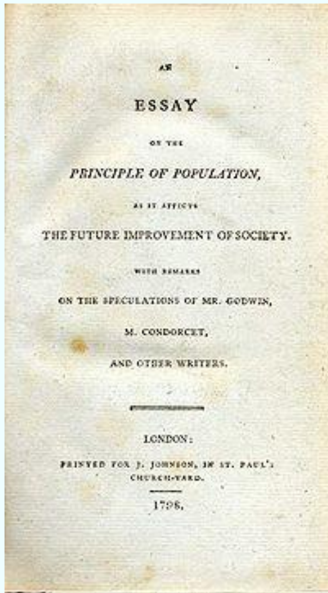
- Three-dimensional assessment of the inflation question
  - Population dynamics
  - ‘Average Error’ in categorised FIDE tournaments
  - Parametric models A(s, c) of Virtual ELO players

    Use of these A(s, c) to assess tournament (players) etc

Summary Results

- Population Analysis
  - the figures do not provide evidence of inflation
  - nor do they disprove the ‘inflation theory’ but ...
    - they do exclude two sources of inflation
- ‘Average Error’ calculations on FIDE-rate tournaments
  - Single-PV analysis picks out ELO-levels of competence
  - show some signs of deflation in the last 20 years
    - i.e. improving standards at ELO Level ‘E’ (for high ‘E’)
- Modelling players using statistical regression:
  - Multi-PV analysis acknowledging most relevant options
  - the ‘optimal parameters’ are reasonably stable over time

1. Population dynamics



AN

ESSAY

ON THE

PRINCIPLE OF POPULATION,

AS IT AFFECTS

THE FUTURE IMPROVEMENT OF SOCIETY.

WITH REMARKS

ON THE SPECULATIONS OF MR. GODWIN,

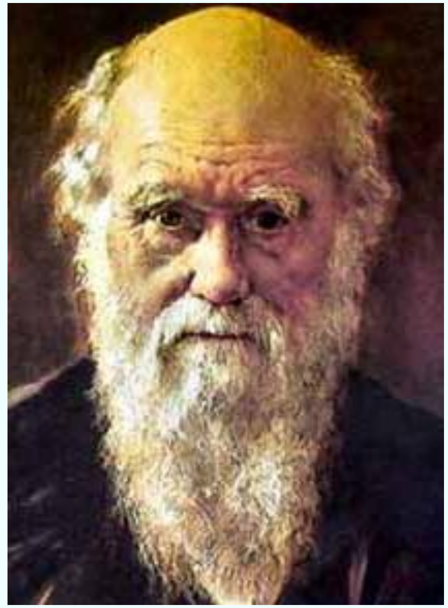
M. CONDORCET,

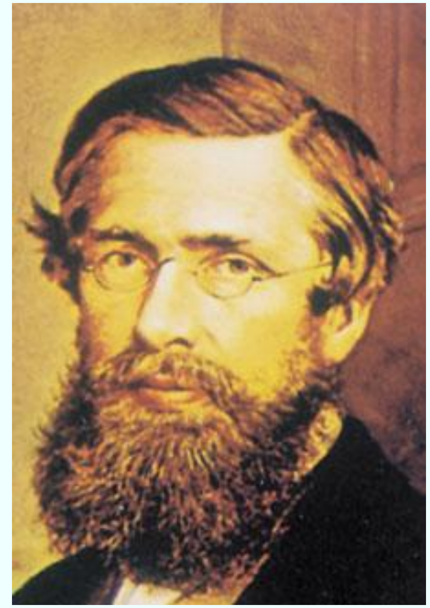
AND OTHER WRITERS.

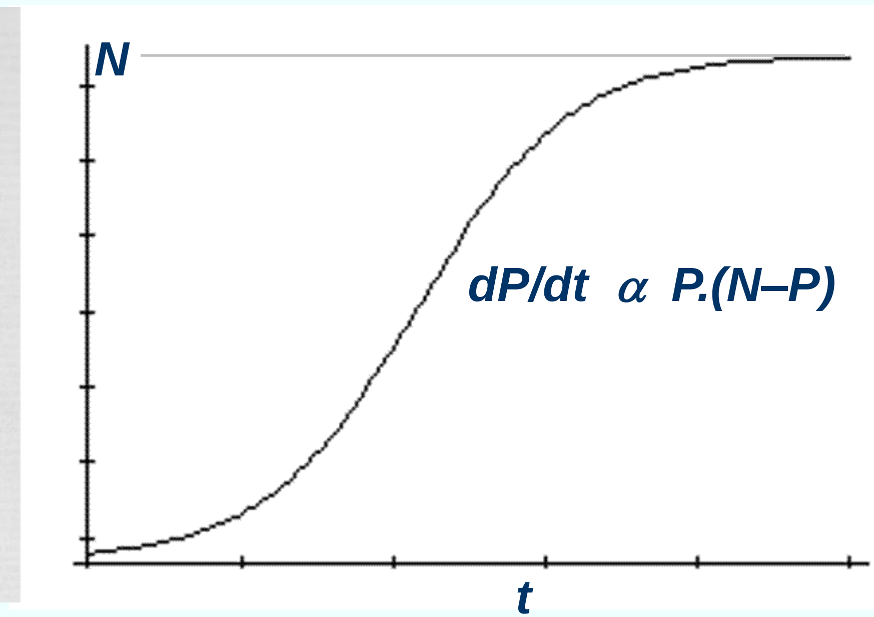
LONDON:

PRINTED FOR J. JOHNSON, IN ST. PAUL'S CHURCH-YARD.

1798.







Malthus (1798):

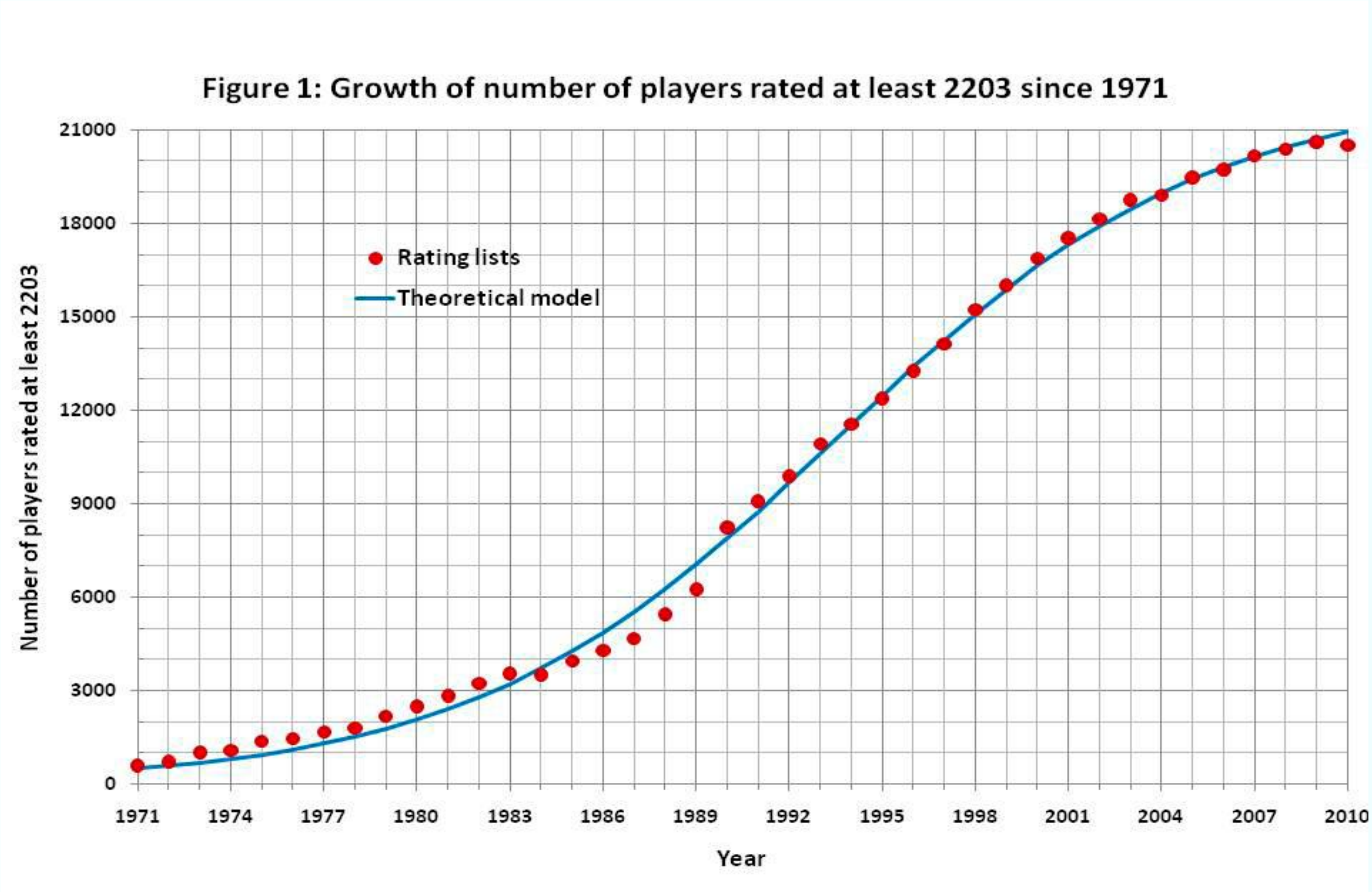
Darwin, Wallace

Verhulst (1838)

Population analysis

- What factors account for the increase in ELO 2203+ players?
  - *Inflation* or other factors
  - Verhulst (1838): $dP/dt \;\alpha\; P.(N\text{–}P) \;\alpha\; P.N - P^2 = a.P - b.P^2$
  - This is the *Logistic Curve*
- the actual data fits well to a Logistic Curve
- The 'fit' supports the idea that:
  - standard population theory explains ELO-population growth
  - the ELO population is not shifting up the scale
  - The ELO population is not expanding up the scale
- ... no support for *ELO Inflation Theory*

Players above ELO 2203 v Logistic Curve



2. Single PV Analysis of Player ‘error’

- RYBKA 3.0 1-cpu run in single-PV mode to 13-ply depth
  - Larry Kaufman estimated depth 14 = 2750
  - We estimate our engine at 2650-2700 (2900 … 2400)
- Run manually in Arena GUI (versions 1.99, 2.01).
  - reproducible except when Rybka ‘stalls’
- All tournaments of category ≥ 11 analysed
  - moves 1-8 ignored; positions > ‘3.00’ ignored
  - 3.77m of 4.00m+ moves analysed
  - two 4-core PCs employed …
- The data is quorate and results seem robust
- Large-scale data needed as benchmark in ‘anti-cheating’ cases

Average Error

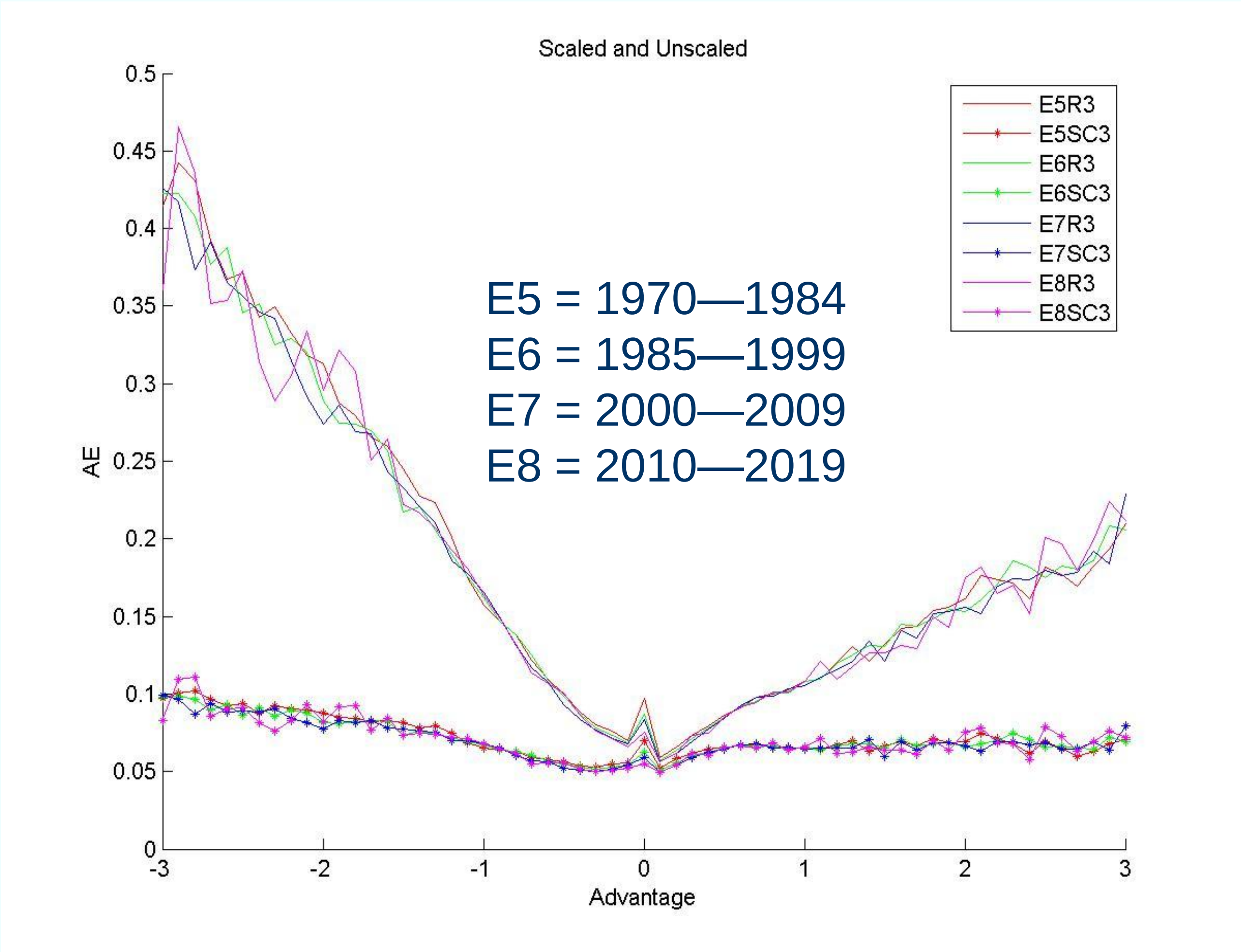
- When played move $\neq$ Rybka's first move,

  error = max(value – value(next position), 0).
- This is logistically simple: perhaps better to use

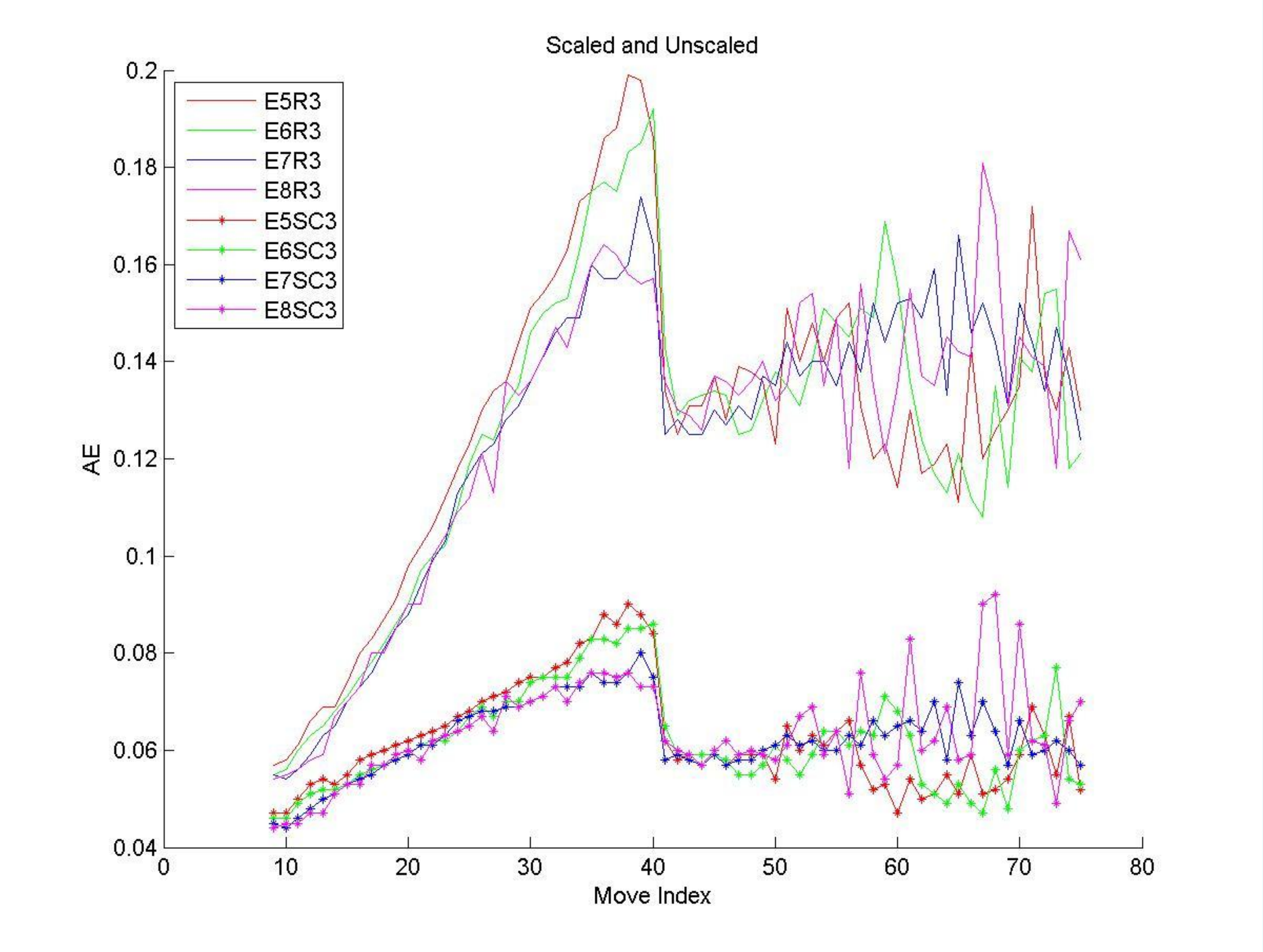
  value(next at depth 12)
- Details differ from Guid/Bratko's work
  - hence we label our errors 'AE' rather than their 'AD'

- A comparison of *Average Error* against *Position Value*
  - larger errors are made in more decisive positions
  - suggests a scaling … 1/(1 + |position value|)

Average Error v Position Value



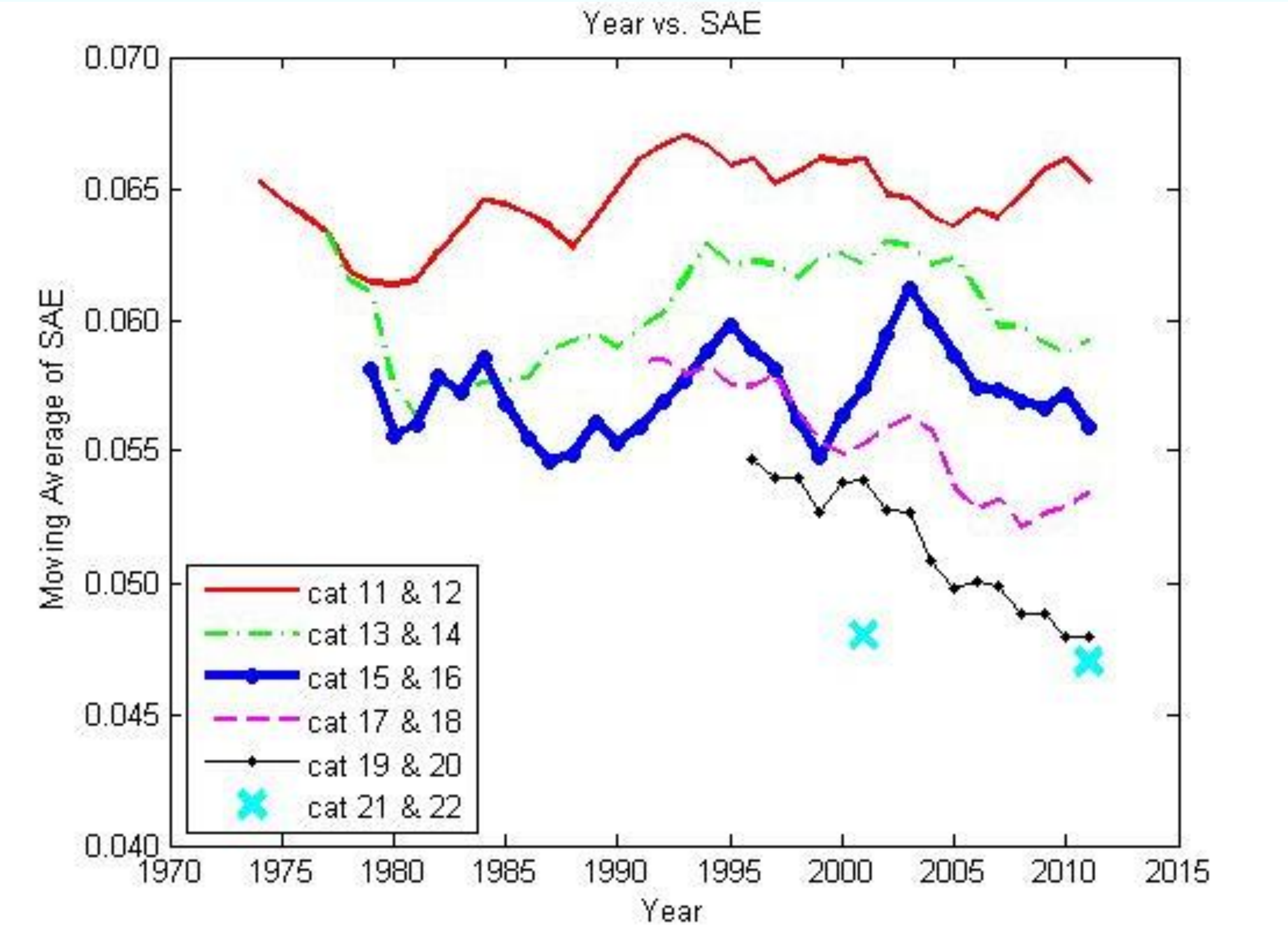
Average Error by Move Number



The effect of time pressure approaching move 40 is clear

Moves 17—32 bridge between opening theory and the worst of Zeitnot

Plot of Scaled Average Error by Category

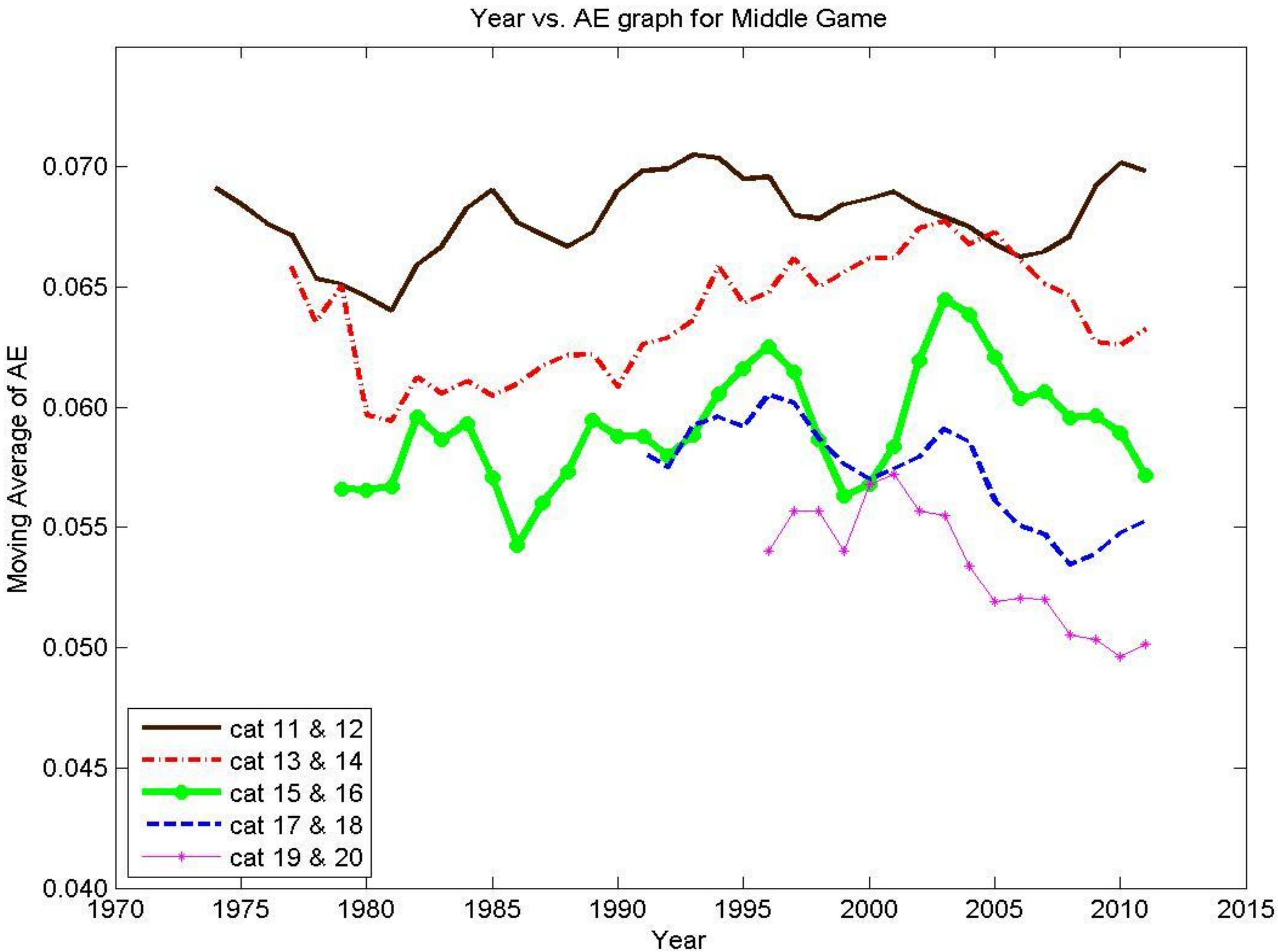


Plot lines would slope up if there were considerable rating inflation.

Some evidence of deflation in higher categories.

Cat 21 and 22, lumped 1996—2001 and 2007—2011.

SAE-by-Category for moves 17 to 32



Curves are similar to case of ‘all moves’; error itself is a little higher.

Overall no-inflation verdict thus independent of today’s greater opening theory knowledge.

3: Parameterised Models of Players

- **Motivation**
  - ***Average Error* does not use the decision's full context**
  - ***Predicting* a player's move requires**

    an 'agent' model of the *fallible player* at their skill level
  - **Hence the need for a range of *Reference Fallible Players***

- **A($\underline{c}$) is an agent with *behaviour parameters* (c_1, c_2, …)**
  - **current model has two parameters:**
    - ***s* ≡ *sensitivity*, *c* ≡ *competence***
- **Prob [A($\underline{c}$) chooses move m_i] α *PF(posval*, v_i, $\underline{c}$)**
- ***s* and *c* determined by statistical regression**

‘Statistically fitting’ agents to human players

- Population used here are ‘Virtual ELO Players’, ELO E
- E = 2700, 2600, 2500 etc
- Virtual players are composite of actual players who ...
  - Are within 10 ELO of, e.g. 2400 and playing a ‘like’ player

- m_0 is the move with the best computer evaluation v_0
- m_i is the *ith* best move and has value $v_i \leq v_{i-1}$
- δ_i is a <u>scaling</u> of $v_i - v_0$
- the probability function PF is defined by:

  $\log(p_i)/\log(p_0) = e\text{^}(-\delta/s)^c$

  this function seems likely to be the best of those defined

Results from defining agents A(*s*, *c*)

- for the Virtual ELO 2400 player, e.g., we define A(*s*, *c*)
- A(*s*, *c*) also has an Average Error AE_c
- thus, we may associate Ae_c with ELO 2400
- now, given a set of players in a tournament ...
- We may determine an A(*s*, *c*) for the tournament

  ... And indeed, for each player in the tournament
- Thus, we may determine a ‘performance ELO’

  for the tournament and each player
- These may be compared with the average FIDE ELO

  for the tournament, and the TPR for each player

A(s, c) results on the training sets

2006—2009 linear interpolation			
Elo	*s*	*c*	IPR
2700±10	.078	.503	2690
2600±10	.092	.523	2611
2500±10	.092	.491	2510
2400±10	.098	.483	2422
2300±10	.108	.475	2293
2200±10	.123	.490	2213

1991—1994 derived IPR values			
Elo	*s*	*c*	IPR
2700±10	.079	.487	2630
2600±10	.092	.533	2639
2500±10	.098	.500	2482
2400±10	.101	.484	2396
2300±10	.116	.480	2237
2200±10	.122	.477	2169

1976—1979 derived IPR values			
2600±10	.094	.543	2647
2500±10	.094	.512	2559
2400±10	.099	.479	2397
2300±10	.121	.502	2277

Inflation would show as IPR > Elo in tables at right. Pretty much none.

Some recent tournaments

Event	cat	Elo	IPR	Diff	Event	cat	Elo	IPR	Diff
Linares 1993	18	2676	2522	-154	Corus 2007	19	2717	2763	+46
Linares 1994	18	2685	2517	-168	Mexico 2007	21	2751	2708	-43
Dortmund 1995	17	2657	2680	+23	Sofia 2007	19	2725	2576	-149
Dortmund 1996	18	2676	2593	-83	Sofia 2008	20	2737	2690	-47
Dortmund 1997	18	2699	2639	-60	Sofia 2009	21	2754	2703	-51
Dortmund 1998	18	2699	2655	-44	Nanjing 2010	21	2766	2748	-18
Dortmund 1999	19	2705	2749	+44	Shanghai 2010	21	2759	2829	+70
Sarajevo 1999	19	2703	2722	+19	Bilbao 2010	22	2789	2904	+115
San Luis 2005	20	2738	2657	-81	Moscow 2010	21	2757	2690	-67
Corus 2006	19	2715	2736	+21	London 2010	20	2725	2668	-57
Sofia 2006	20	2744	2744	0	**Averages**	**19**	**2722**	**2690**	**-32.6**

IPRs are reasonable; half of shortfall is from Linares 1993-94.

No support for inflation hypothesis here either.

The Canadian Open, July 9-17, 2011

- **9 round Swiss: 149 players (115 with FIDE ratings)**
- **623 games available and analysed (of 647 played)**

Whole event	CanR	TPR	IPR		Restrict	CanR	FIDE	IPR
Average	2144	2142	**2117**		to 115	2211	2139	**2203**
St. Deviation	258	261	**379**		FIDE-	229	220	**345**
Wtd. by games	2156	2154	**2134**		rated	2221	2147	**2219**
Wtd. by moves	2173	2172	**2161**		players:	2236	2161	**2242**

- **Can compare IPRs with TPRs and with FIDE ELO ratings**
- **Impression is that Canadian players here are too low in FIDE ELO**

Conclusions

- Three-dimensional assessment of the *ELO Inflation* issue
- Population analysis does not support *inflation theory*
- Average Error hints at *deflation* rather than *Inflation*
- Multi-PV analysis is effective on a smaller scale
  - yields credible *Intrinsic Performance Ratings*
  - these IPRs correlate well with ELO
  - ... a vote of confidence for both

The Way Ahead ... Some thoughts

- Improved infrastructure for our computation experiments
  - *repeatability* requires engines which do not *stall*
  - a database to store engine-evaluations of positions
  - automated exploitation of distributed computing resources
- Integration of two statistical approaches
  - Statistical regression
  - Bayesian inference
- Further exploitation
  - Our analyses can be cloud-sourced in real-time
  - Application to other *Fallible Decision Maker* areas

Our thanks to ...

- The ARENA GUI programmers for full-analysis scripting
- Programmers associated with TOGA II and RYBKA
- Univ. Of Buffalo CSE and Univ. of Montreal for support
- Tamal Biswas, managing the data, creating the graphs
- Hugh Brodie and David Cohen for the Can. Open games