A Framework for Generating Data to Simulate Application Scoring

Credit Scoring and Credit Control XII,
University of Edinburgh,

Kenneth Kennedy,
Dublin Institute of Technology, Ireland

Contents

- Artificial Data
  - Motivation
- Methodology
  - Data Generation
  - Label Application
- Population Drift
- Conclusion

Data Sources

- Benchmark studies
  - West (2000) 302 citations
  - Baesens et al. (2003) 190 citations
- Proposed method
- Concerns
  - Superior performance
  - Data

UCI Datasets

- UCI Machine Learning Repository
  - Allows comparison of techniques
  - Over-reliance
    - Overfitting (Drummond, 2006)
    - Data bias (Keogh & Kasetty, 2003)
      - *"the conscious or unconscious use of a particular set of testing data to confirm a desired finding."*
  - Ignore conditions
  - Not representative (Saitta & Neri, 1998)
    - Size and distribution
    - Characteristics

Real World Datasets

- Real Data
  - Gold standard
    - *"Close engagement with the data, owners of the data, and problems of the data is vital"*. (Hand, 2010)
  - Benefits of data sharing
  - Data sharing culture (Fischer & Zigmond, 2010)
  - Negative career impact
  - Limited resources
  - Property rights and Legal issues

Artificial Data

- Manipulation of various parameters used in the evaluation process
  - Design specific experiments,
  - Under certain conditions of interest in a relatively precise manner (Scott & Wilkins, 1999; Japkowicz and Shah, 2011).
- Advance some research proposed direction in tool development

Data Sources

- Irish mortgage market in 2007
- Sources
  - Demographic data from the Central Statistics Office (CSO), Ireland (2010)
  - Housing statistics published by the Department of Environment, Heritage and Local Government (Irish Gov.) (2008)
  - Central Bank of Ireland technical report (McCarthy and Quinn, 2010)
  - Moody's research report on why Irish borrowers default (2010)
  - Credit risk expert

Methodology

Generate Instances – Borrower Features

Feature	Value	Source
First-Time-Buyer	1,0	Irish Gov.
Age Group	18-25, 26-30, 31-35,...46-55	Irish Gov.
Income Group	30-40k, 40k-60k, 60k-80k,…	Irish Gov. / Moody's
Employment Sector	Health, Hospitality, Construction…	CSO
Occupation	Manager, Employee, Trade…	Irish Gov.
Household Compos.	1 Adult, No Children < 18;…	CSO
Education	Primary,…,3rd Level Higher Degree	CSO
Expenses-to-Income	Ratio	CSO

Generate Instances – Loan Features

Feature	Value	Source
Location	Dublin, Cork, Galway,…	Fitches
New Home	1,0	Irish Gov.
Loan Value	50k-100k, 100k-150k,…450k-900k	Irish Gov. / Moody's
LTV	45%,55%,60%,…,97.5%,100%	Irish Gov.
Loan Term	20, 25, 30, 35, 40	Irish Gov.
Loan Rate	Fixed, Variable, Tracker	Irish Gov.
House Value	Loan Value * LTV	Irish Gov. / Moody's
MRTI	Ratio	-

Generate Instances – Prior Probabilities

Location	Prior Probability
Dublin	32%
Cork	15%
Galway	7%
Limerick	4%
Waterford	3%
Other	39%

New Home	Prior Probability
1	46%
0	54%

Generate Instances – Conditional Prior Probabilities

FTB	Location	New Home	Conditional Prior Probability
1	Dublin	1	0.41
0	Dublin	1	0.59
1	Dublin	0	0.30
0	Dublin	0	0.70
1	Cork	1	0.38
…	…	…	…

Methodology

Calculate Risk Score - Feature Risk

- Assess risk of feature
  - assign one of 7 monotonic levels

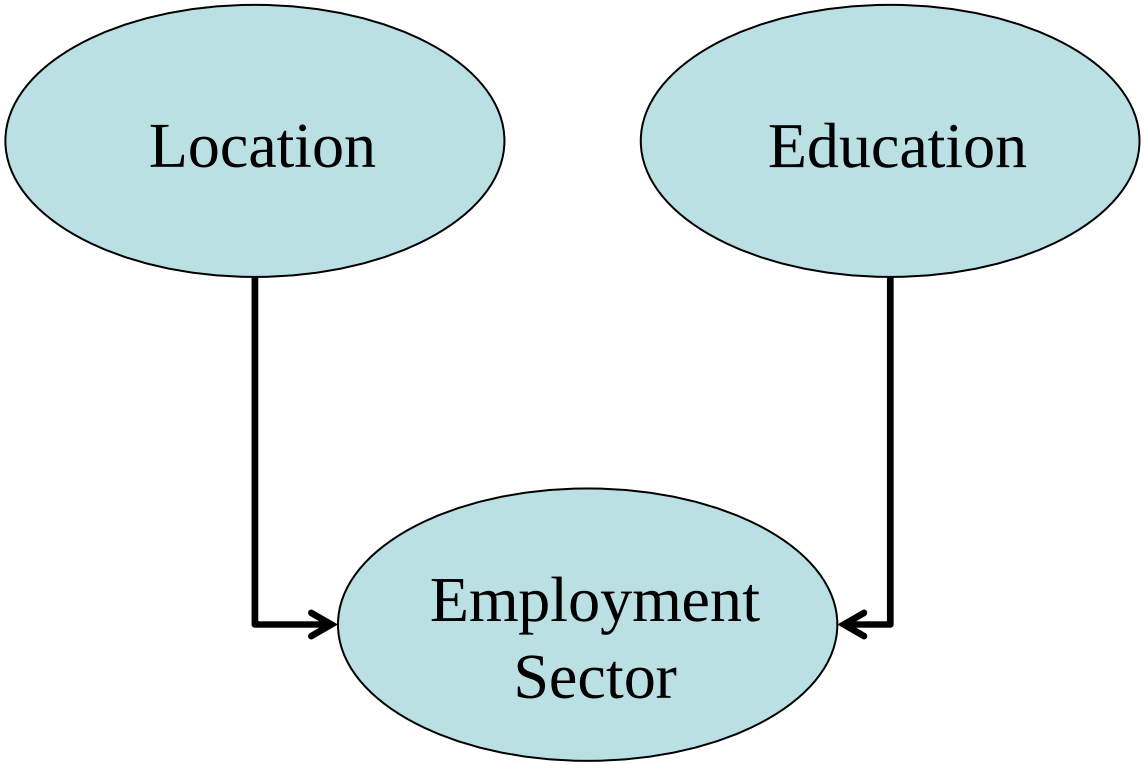
Feature	Value	Feature Score
Location	Dublin	2
FTB	1	6
Age	26-30	5
Education	3rd Level Non-Degree	3
Employment Sector	Construction	6
…	…	…

Calculate Risk Score - Feature Risk

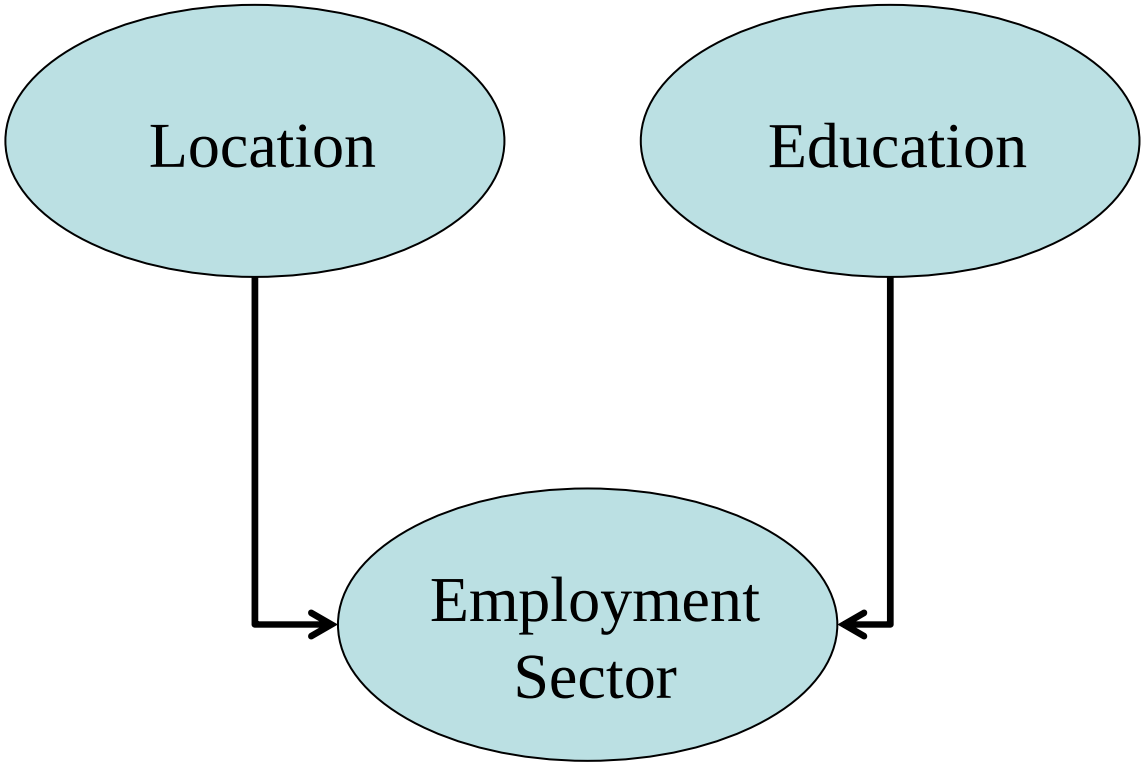
- Assess risk of feature
  - assign one of 7 monotonic levels

Feature	Value	Feature Score
Location	Dublin	2
FTB	1	6
Age	26-30	5
Education	3rd Level Non-Degree	3
Employment Sector	Construction	6
…	…	…

Calculate Risk Score - Interactions

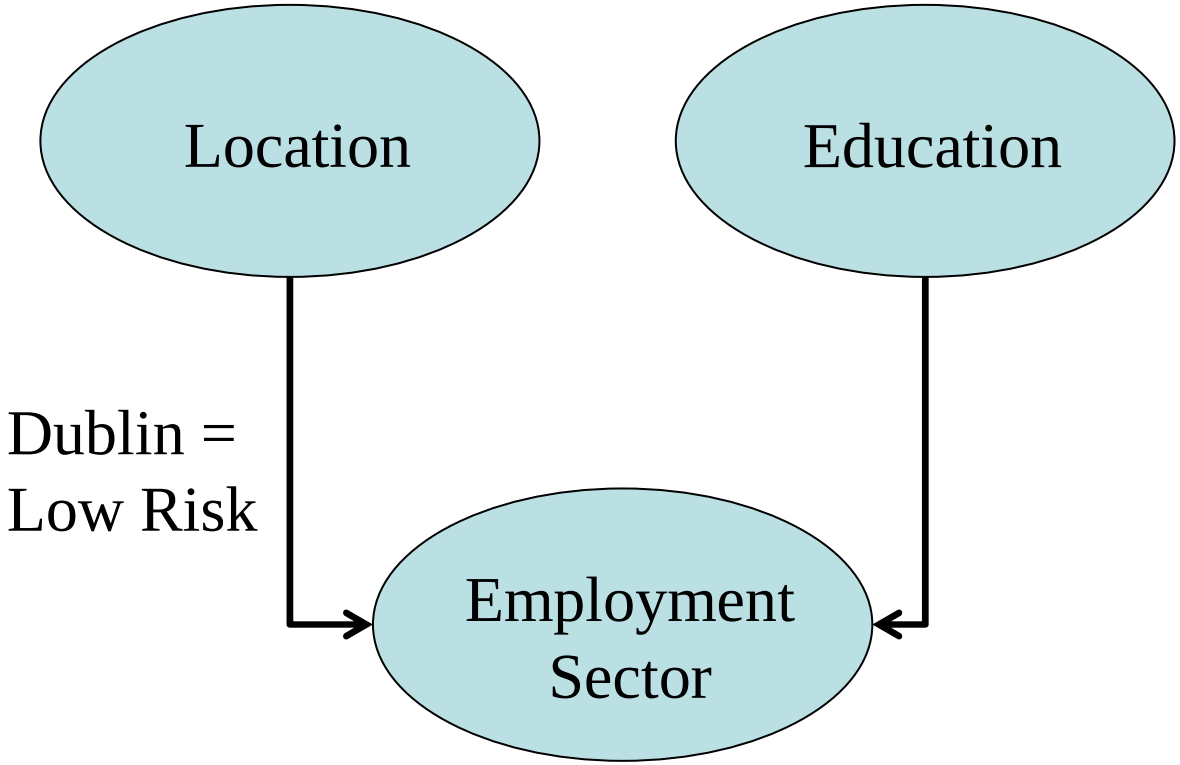


Calculate Risk Score - Interactions

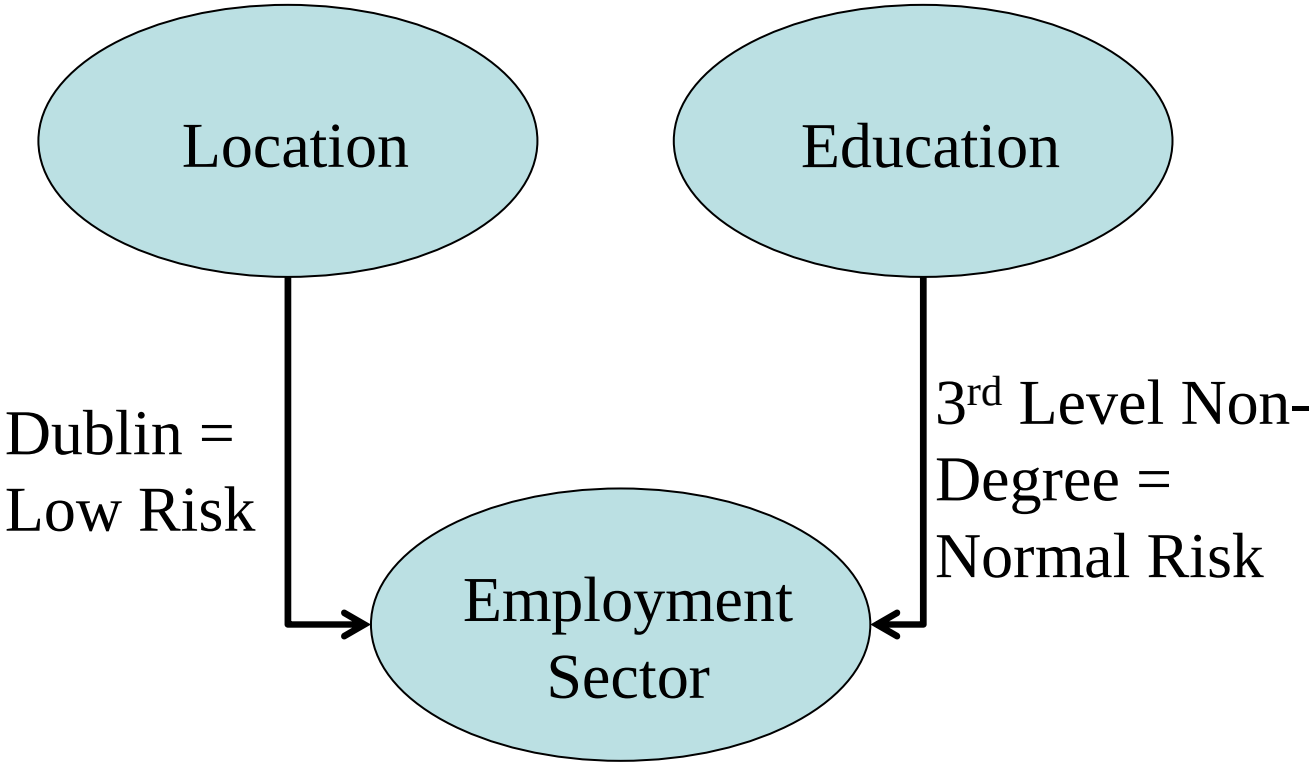


Construction = High Risk
(5,6,7)

Calculate Risk Score - Interactions

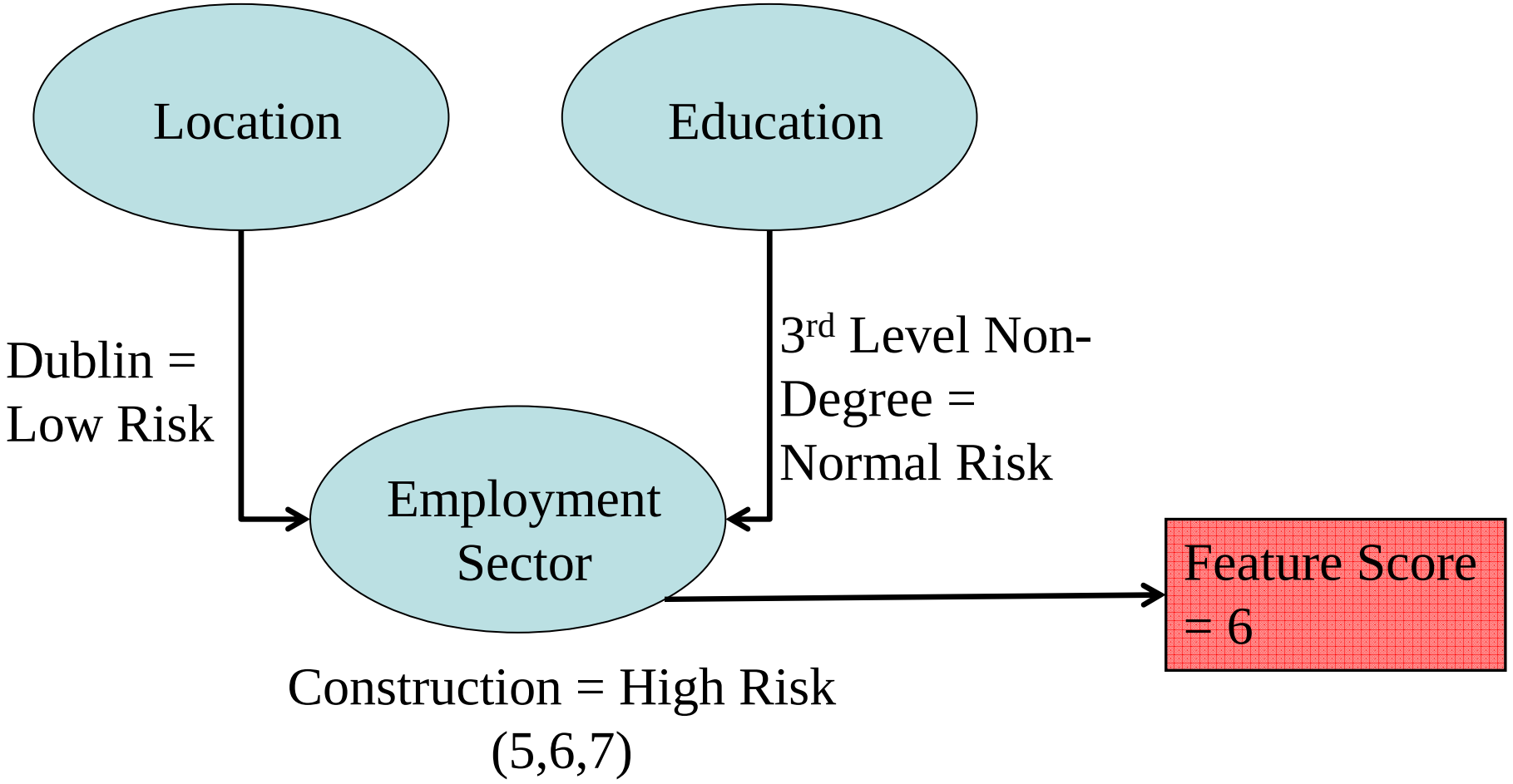


Calculate Risk Score - Interactions



Construction = High Risk (5,6,7)

Calculate Risk Score - Interactions



Calculate Risk Score

-

- Accumulate Feature Scores
- Add noise

Feature	Value	Feature Score
Location	Dublin	3
FTB	1	6
Age	26-30	5
Education	3rd Level Non-Degree	3
Employment Sector	Construction	6
…	…	…
Risk Score	-	**76**

Methodology

Apply Labels

Risk Score	Label
90.15	Bad
90.00	Bad
89.70	Bad
88.60	Bad
87.00	Bad
86.50	Good
86.40	Good
86.5	Good
86.5	Good
…	Good
22.50	Good
21.70	Good

- Default rate, e.g. 2.5%
- Swap Good -> Bad

Apply Labels

Risk Score	Label
90.15	Bad
90.00	Bad
89.70	Bad
88.60	Bad
87.00	Bad
86.50	Good
86.40	Good
86.5	Good
86.5	Good
…	Good
22.50	Bad
21.70	Good

- Default rate, e.g. 2.5%
- Swap Good -> Bad

Comparison

Dataset	TNR	TPR	Average Class Accuracy
Australia	0.876	0.876	0.876
German	0.717	0.720	0.719
Artificial Dataset	0.808	0.926	0.867

Population Drift

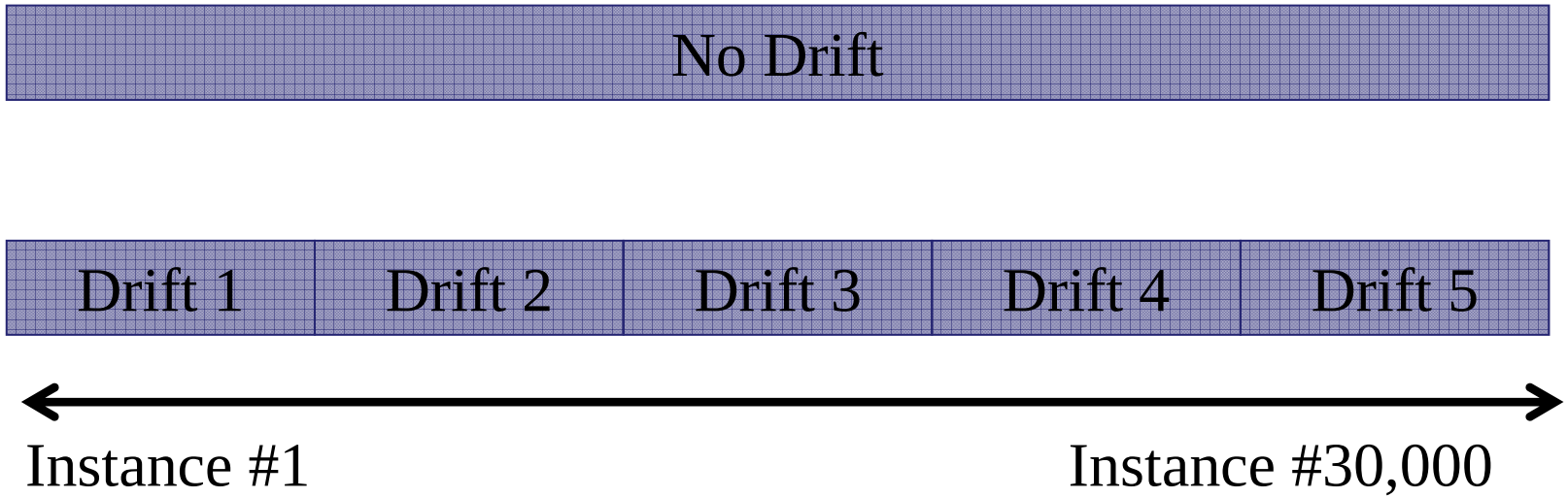
- Demographic change
- Marketing campaigns (own/competitor)
- Errors in coding, data capture, human input
- Adaptive customer behaviour
- Variability of the economic environment
- Performance window

Population Drift

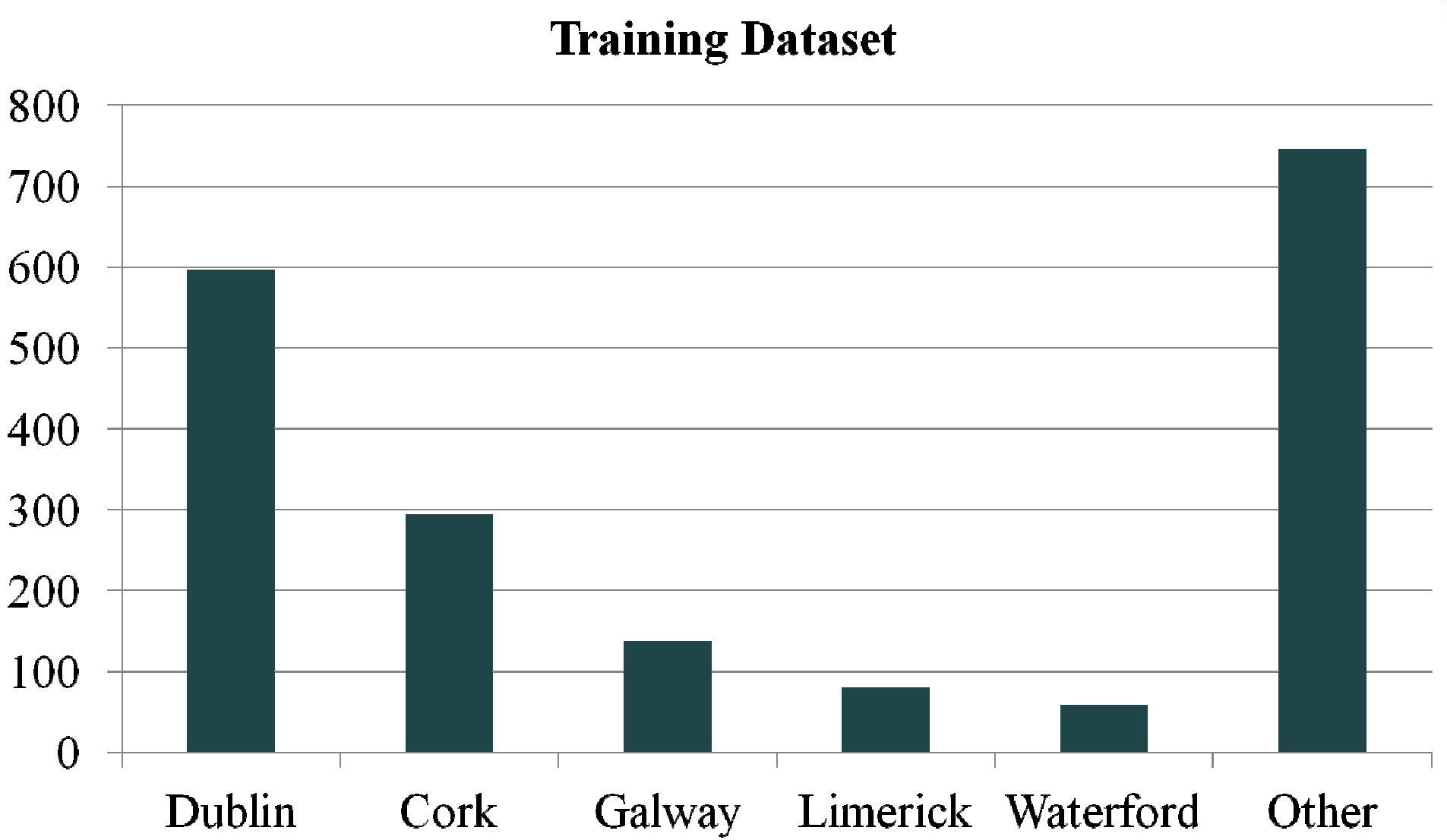
- Training data
  - 2000 instances
  - Remove unrealistic: 1912 instances
  - Default rate 2.5% (48 instances)
  - Noise with 0.5 standard deviation
  - Record cut-off risk score
  - Swap rate 0.33% (6 instances)
  - 70:30

Population Drift

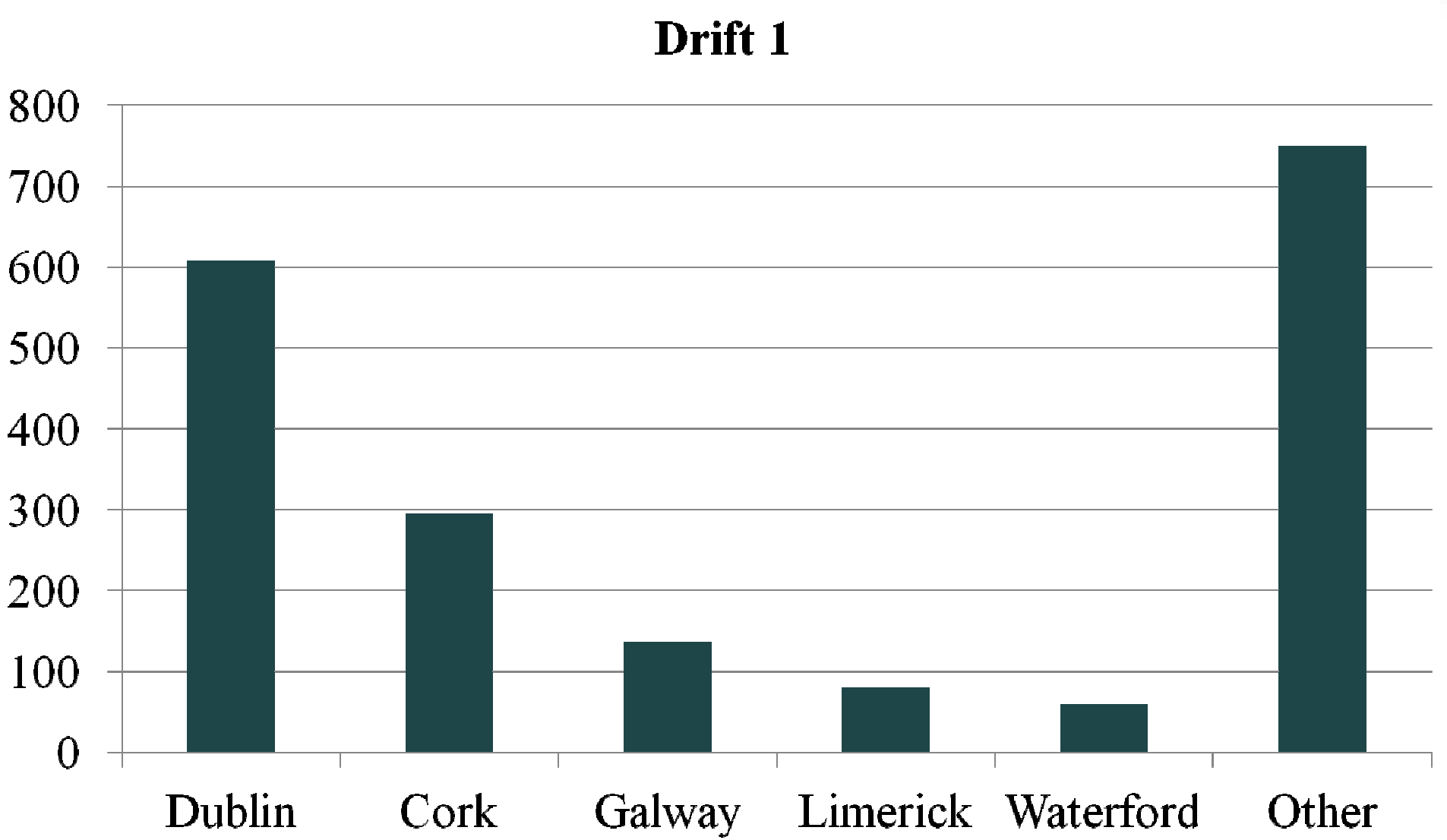
- Test data



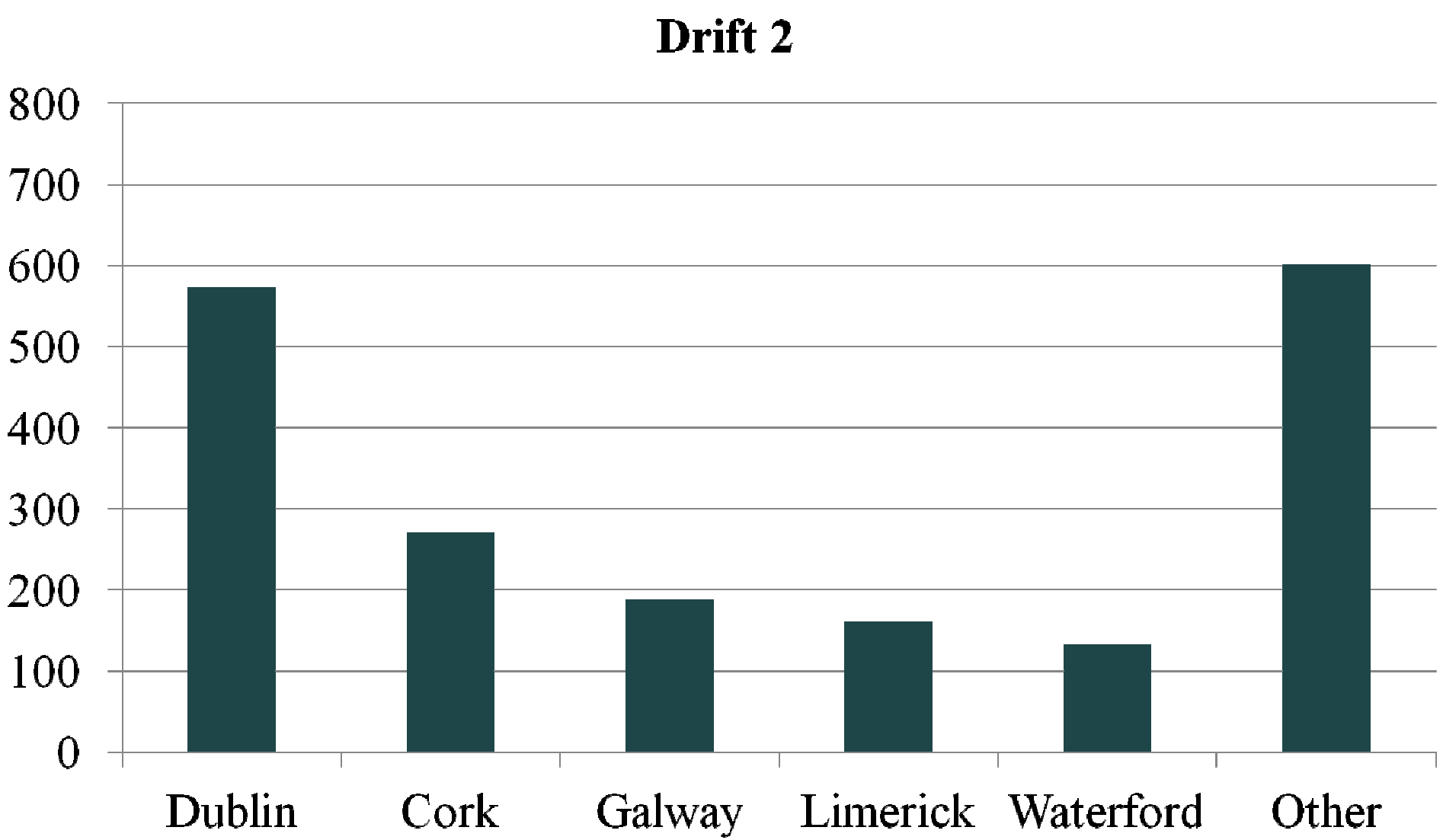
Population Drift



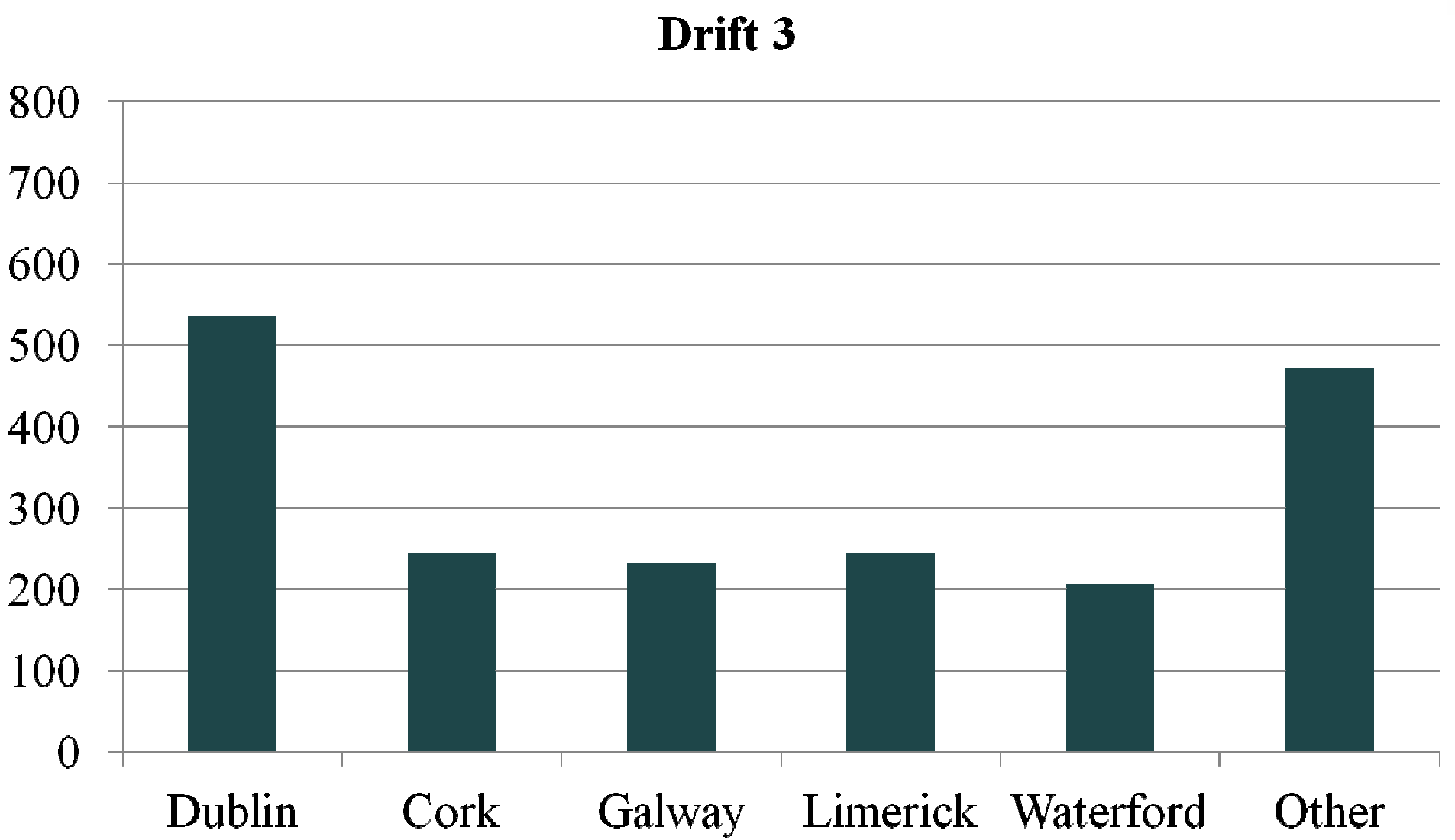
Population Drift



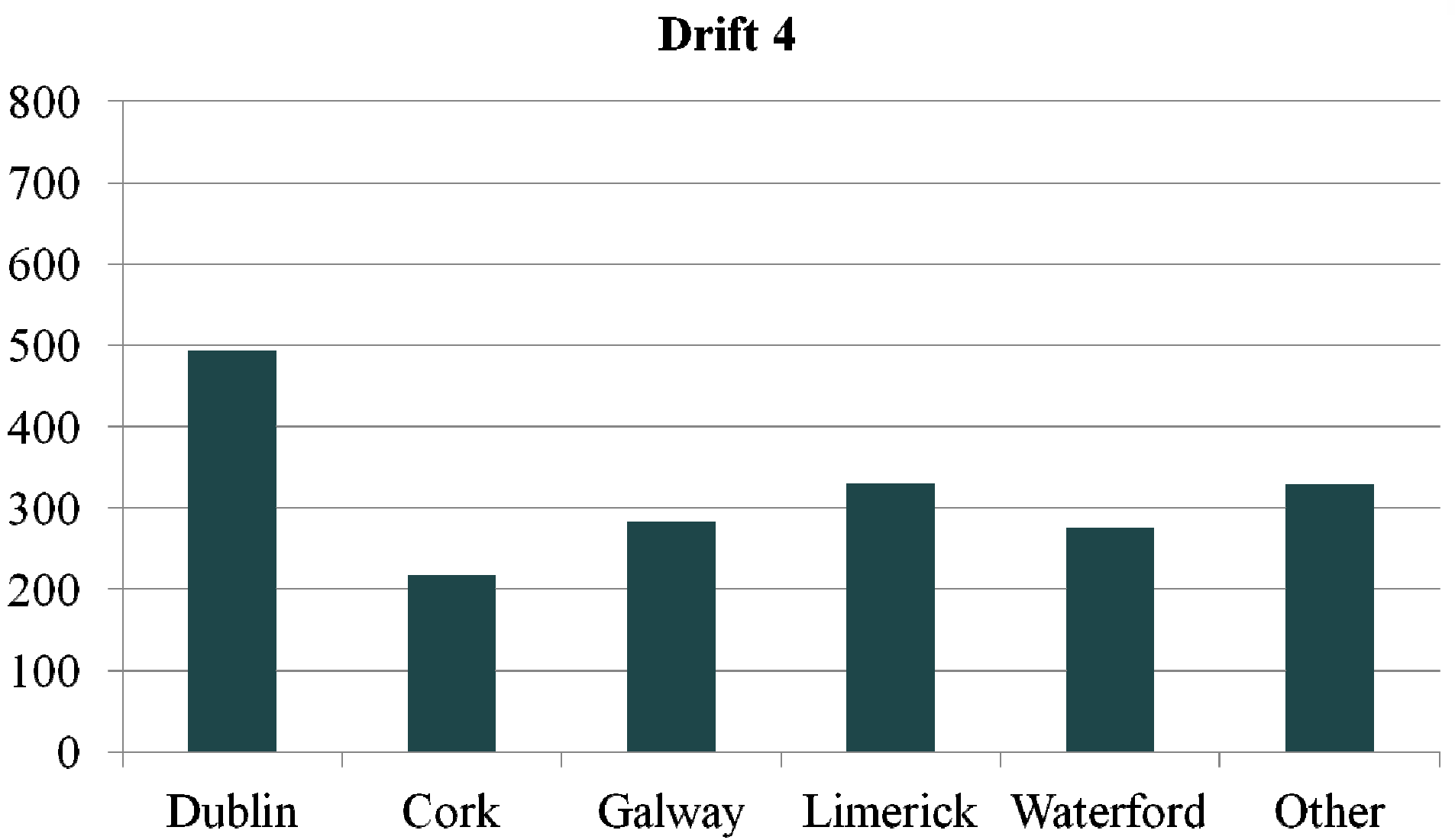
Population Drift



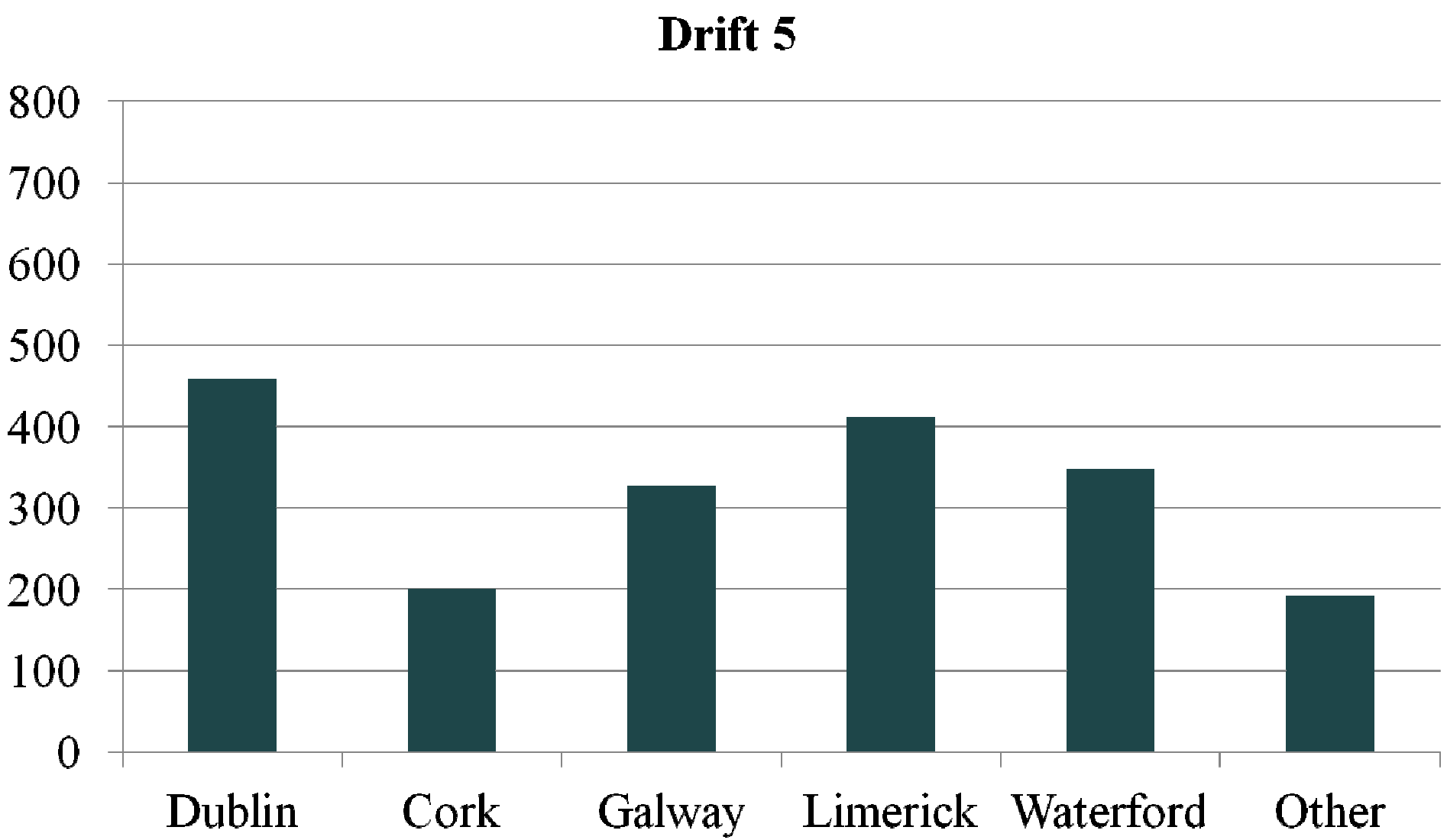
Population Drift



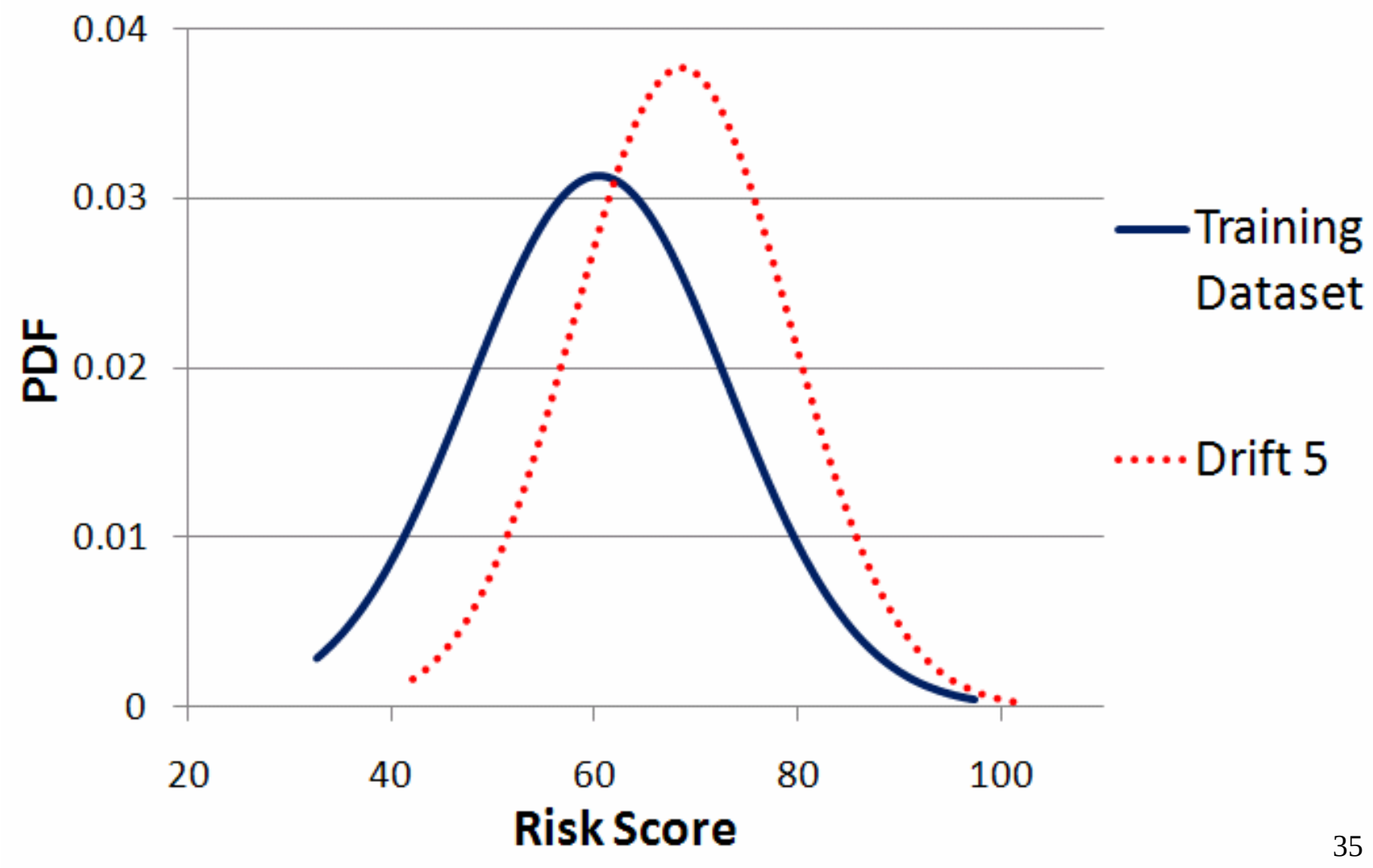
Population Drift



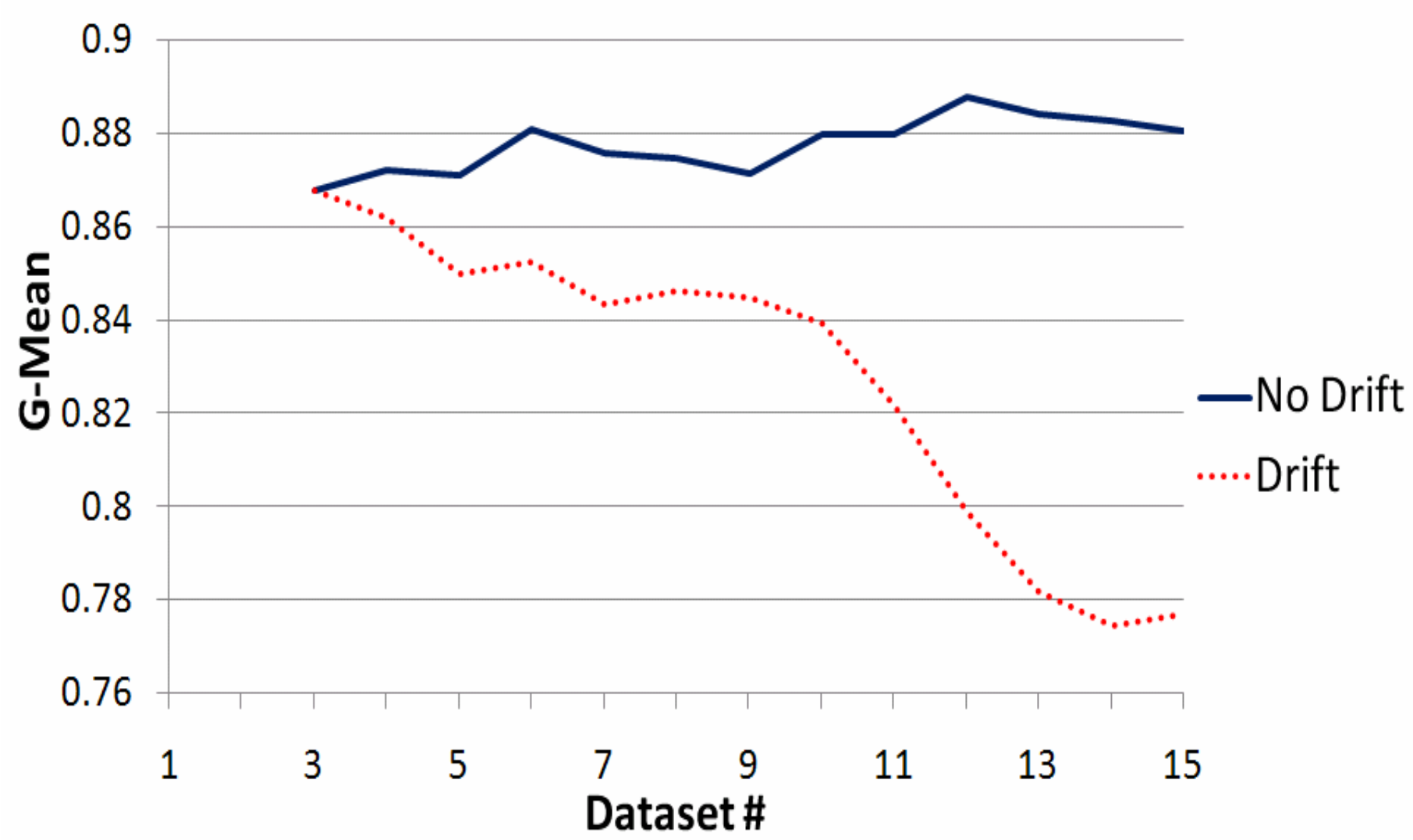
Population Drift



Population Drift



Population Drift



Conclusion

- Impediments to real world datasets
- UCI
- Artificial data
  - Generate
  - Label
  - Limitations
  - Direction
- Population drift
- kennedykenneth@gmail.com

References

- Alaiz-Rodrıguez, R., and Japkowicz, N. (2008). Assessing the impact of changing environments on classifier performance. In Proceedings of the Canadian Society for computational studies of intelligence, 21st conference on Advances in artificial intelligence (pp. 13–24). Springer-Verlag.
- Central Statistics Office, Ireland (2008). Statistical Yearbook of Ireland, 2008 Edition. http://www.cso.ie/releasespublications/statistical_yearbook_ireland_2008.htm. Accessed 3rd February 2011.
- Moody's Global Credit Research (2010). What Drives Irish Mortgage Borrowers to Default. http://www.alacrastore.com/ research/moodys-global-credit-research-What_Drives_Irish_Mortgage_Borrowers_To_Default-PBS_SF226391. Accessed 3rd February 2011.
- Department of the Environment, Heritage and Local Government, Ireland (2008). Latest House Prices, Loans and Profile of Borrowers Statistics. http://www.environ.ie/en/Publications/StatisticsandRegularPublications/HousingStatistics/. Accessed 3rd February 2011.
- Fischer, B., and Zigmond, M. (2010). The essential nature of sharing in science. Science and engineering ethics, (pp. 1–17).

References

- Baesens, B., Gestel, T. V., Viaene, S., Stepanova, M., Suykens, J., and Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. J Opl Res Soc, 54, 627–635
- Hand, D. (2010). Fraud Detection in Telecommunications and Banking: Discussion of Becker, Volinsky, and Wilks (2010) and Sudjianto et al.(2010). Technometrics, 52, 34–38.