

# **Exploratory Analysis of Massive Data for Distribution Fault Diagnosis in Smart Grids**

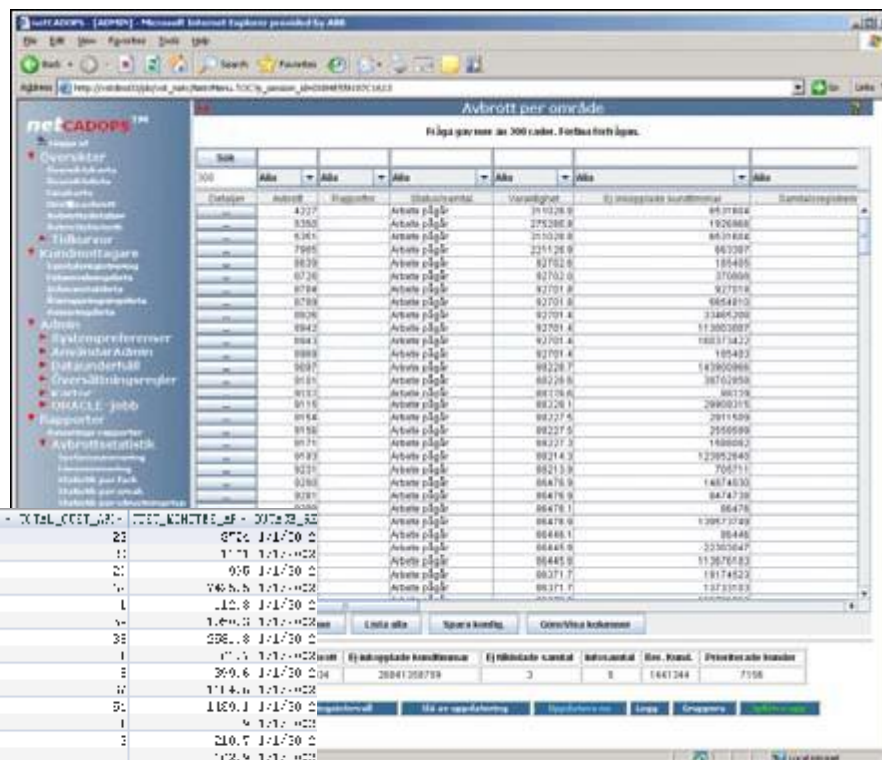
Yixin Cai, Mo-Yuen Chow  
Electrical and Computer Engineering,  
North Carolina State University  
July 2009

- Introduction
- Integrate data
- Evaluate a single feature
- Build a fault cause classifier
- Summary

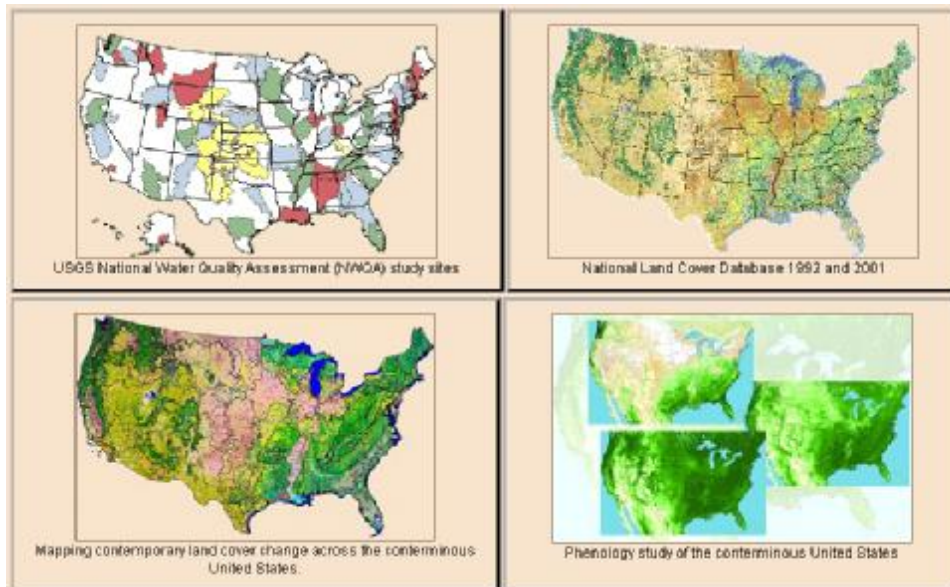
- Intelligent fault management
  - detection, recording, location, diagnosis, restoration, ...
- Problem of interest
  - Diagnosis: predict what the root cause is based on the available information before the engineers go on-site
  - Help (not replace) the engineers to identify the root cause faster
- Challenges
  - Stochastic nature of faults
  - Noisy data with errors
  - More and more incoming data in Smart Grids

- Data sources
  - Utility OMS database

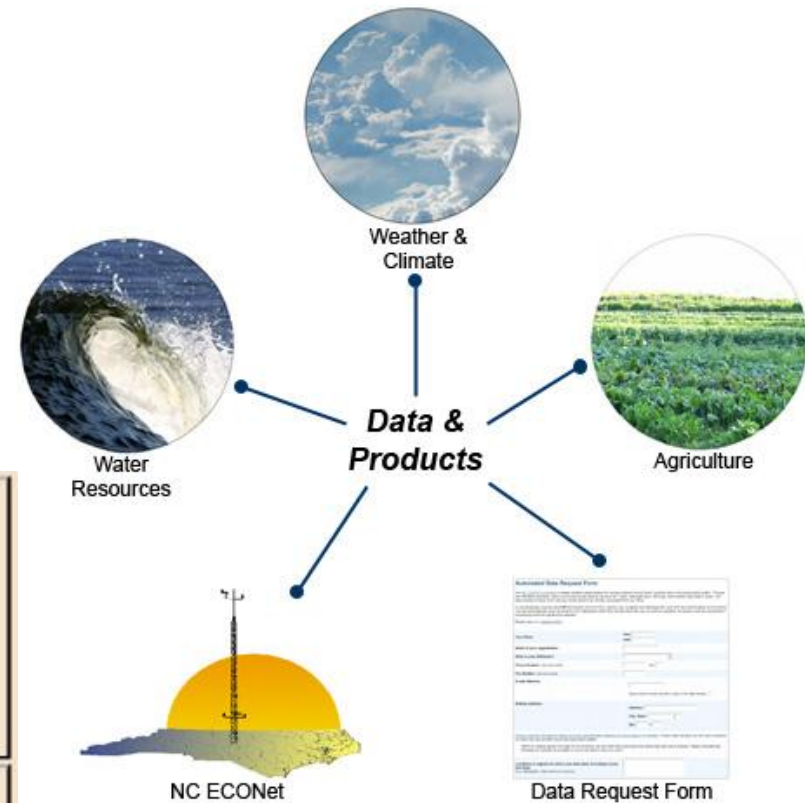
VTID#_E3I -	OFFICE_CD -	OFFICE_NAME -	SUBSTATION -	FEEDER_CD -	FEEDER_NAME -	HOW_EVS_CD -	HOW_EVS_NAME -	PHASED -	VTID_CODE_NAME -	VTID_CODE_DESCRIPTION -
090500	304	Lake City	70300	E02	T010002	01	WATER CO WILSONS BOND	20	0100	1-1-70 0
010501	304	Lake City	70300	E01	T010002	01	WATER CO WILSONS BOND	10	1000	1-1-70 0
090500	310	Wilmington	70700	E07	T010007	01	WILMINGTON WATERWORKS	20	0100	1-1-70 0
010502	400	Wilmington	70400	E14	T010007	01	WILMINGTON WATERWORKS	14	1400	1-1-70 0
090508	100	Charlotte	70100	E11	T010011	01	CANTON W. JACOBSON STKS	11	1100	1-1-70 0
010509	100	Charlotte	70100	E05	T010011	01	CANTON W. JACOBSON STKS	05	1000	1-1-70 0
090506	104	Charlotte	70304	E04	T010011	01	EDDIE GREEN EDDIE GREEN	04	0400	1-1-70 0
010506	104	Charlotte	70304	E01	T010011	01	EDDIE GREEN EDDIE GREEN	01	1000	1-1-70 0
090507	300	Charlotte	70300	E05	T010010	01	SCATTER NORTHMONTA WASH	05	3000	1-1-70 0
010501	300	Charlotte	70300	E01	T010010	01	SCATTER NORTHMONTA WASH	01	1000	1-1-70 0
090502	200	Charlotte	70100	E03	T010010	01	STANFORD EXETERSTON STKS	03	1000	1-1-70 0
010502	200	Charlotte	70100	E01	T010010	01	STANFORD EXETERSTON STKS	01	1000	1-1-70 0
090500	301	Charlotte	70300	E01	T010010	01	ELLIOTT DOUGLASS DEBY	01	2100	1-1-70 0
010505	301	Charlotte	70300	E07	T010010	01	ELLIOTT DOUGLASS DEBY	07	1000	1-1-70 0
090500	100	Charlotte	70700	E11	T010011	01	LEWIS LARRY REMBLE STKS	11	0200	1-1-70 0
010505	100	Charlotte	70700	E14	T010011	01	LEWIS LARRY REMBLE STKS	14	1000	1-1-70 0
090500	101	Charlotte	70300	E01	T010011	01	GREENBUSH * GREENBUSH	01	0000	1-1-70 0
010505	101	Charlotte	70300	E04	T010011	01	GREENBUSH * GREENBUSH	04	1000	1-1-70 0
090502	400	Charlotte	70300	E04	T010011	01	HARTFORD J. MAIF STREET	04	0000	1-1-70 0
010505	400	Charlotte	70300	E01	T010011	01	HARTFORD J. MAIF STREET	01	1000	1-1-70 0
090506	106	Charlotte	70300	E03	T010010	01	ROXBORO EXETERSTON STKS	03	3000	1-1-70 0
010506	106	Charlotte	70300	E01	T010010	01	ROXBORO EXETERSTON STKS	01	1000	1-1-70 0
090500	400	Charlotte	70300	E03	T010010	01	EMILY * WILSON PERE ORMS	03	3000	1-1-70 0
010501	400	Charlotte	70300	E01	T010010	01	EMILY * WILSON PERE ORMS	01	1000	1-1-70 0
090502	300	Charlotte	70200	E01	T010010	01	MATCH LINDMONT STKS	01	6000	1-1-70 0
010502	300	Charlotte	70200	E04	T010010	01	MATCH LINDMONT STKS	04	1000	1-1-70 0
090504	301	Charlotte	70600	E01	T410001	01	SALFORD WILSONSTON STKS	01	4000	1-1-70 0
010508	301	Charlotte	70600	E07	T410001	01	SALFORD WILSONSTON STKS	07	1000	1-1-70 0
090504	400	Charlotte	70300	E01	T010011	01	ELM * WILSONSTON STKS	01	0200	1-1-70 0
010505	400	Charlotte	70300	E04	T010011	01	ELM * WILSONSTON STKS	04	1000	1-1-70 0



- Data sources
  - Utility OMS database
  - Public database on weather, environment, geographic features, etc.



<http://landcover.usgs.gov/usgslandcover.php>



<http://www.nc-climate.ncsu.edu/products>

- Data sources
  - Utility OMS database
  - Public database on weather, environment, geographic features, etc.
  - Private vendors

**PRODUCTS**    **SERVICES**    **SAMPLES**    **NEWS & EVENTS**    **WHAT'S NEW**    **ABOUT US**

**WELCOME TO MAPMART**  
Global Mapping Solutions

IntraSearch, Inc., a Colorado corporation since 1951, is a fully integrated professional mapping services and product distribution company comprised of three components: **IntraSearch Mapping**, **MapMart**, and **SportsMapping**. The MapMart division distributes satellite imagery, aerial photography, and numerous map products via its e-commerce site, **MapMart.com™**. IntraSearch's **SportsMapping™** division provides map products and visualization services to the sports industry.

**CUSTOM MAPPING SERVICES**

IntraSearch has in-house mapping services staffed with experienced personnel to meet all engineering, planning and GIS mapping needs:

- [Aerial Photography](#)
- [Orthophotography](#)
- [Photographic Lab Services](#)
- [Topographic Mapping](#)
- [Environmental Mapping](#)
- [Golf Course Mapping](#)
- [MapMart Data Hosting Services](#)
- [More...](#)

**PRODUCTS**

**DATA STORE**

Click on a window below to access customizable digital data and imagery sources for the US and around the world. Many are available to download within minutes of purchase.

aerial photography	satellite imagery	digital elevations	topographical maps
digital street maps	digital vector maps	demographic data	world data by country

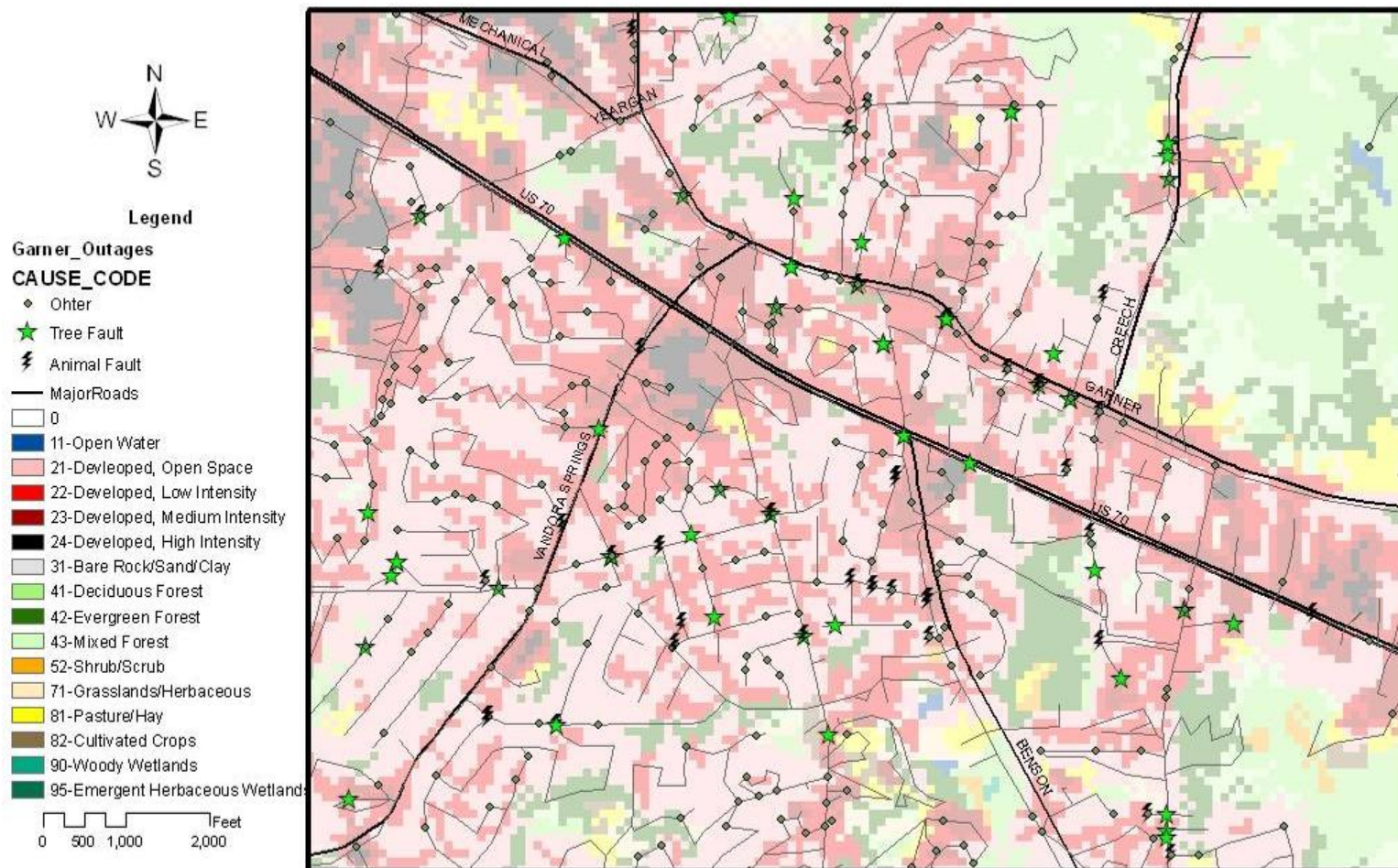
**SOFTWARE STORE**

Click on a window below to access customizable digital data and imagery sources for the US and around the world. Many are available to download within minutes of purchase.

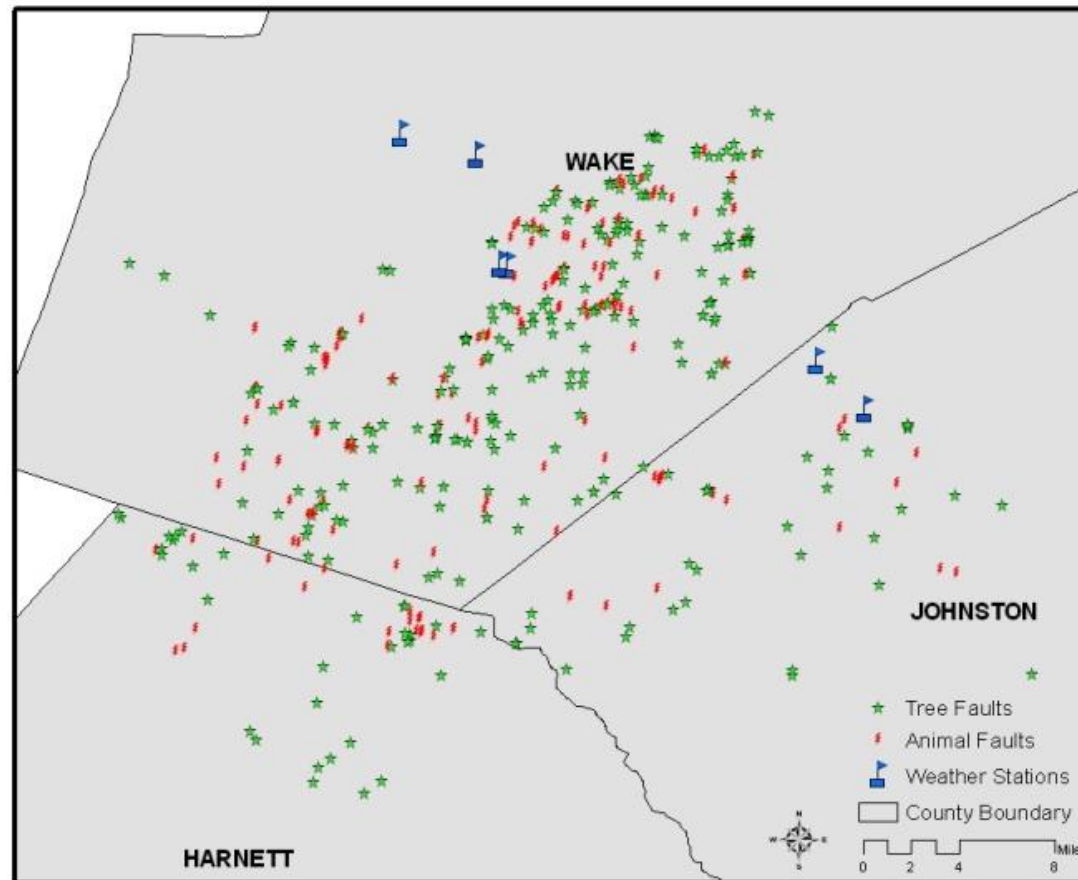
MapInfo	ESRI	Global Mapper	LizardTech
Pictometry			

<http://www.mapmart.com/Default.aspx>

- Data integration under GIS framework
  - Spatial relation



- Data integration under GIS framework
  - Spatial relation
  - Spatial-temporal relation

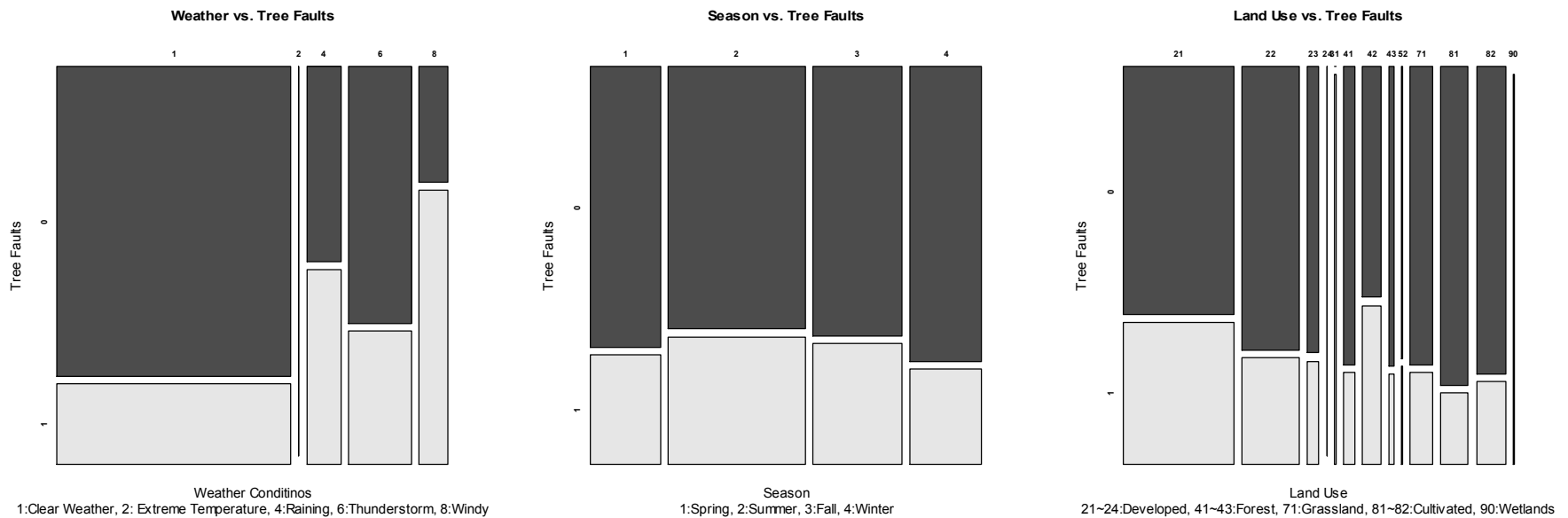




- Data preprocess
  - Define the root cause of interest
  - Clean errors and noises
  - Extract features

- Categorical features

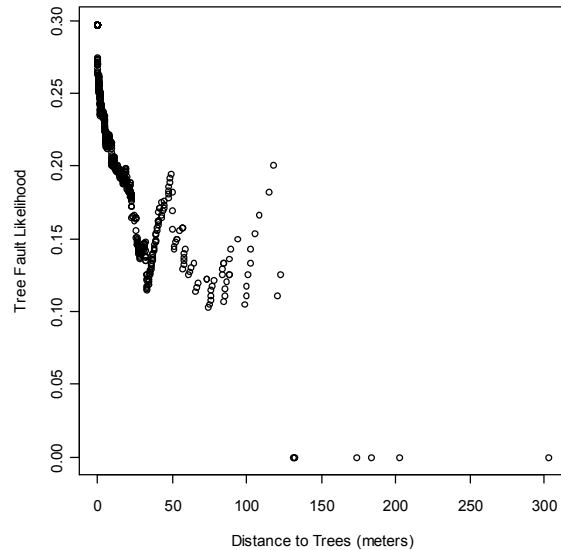
- Likelihood measure  $L_{i,j} = P(o_i | X = x_j) = \frac{N_{i,j}}{N_j}$
- (Mosaic) plot



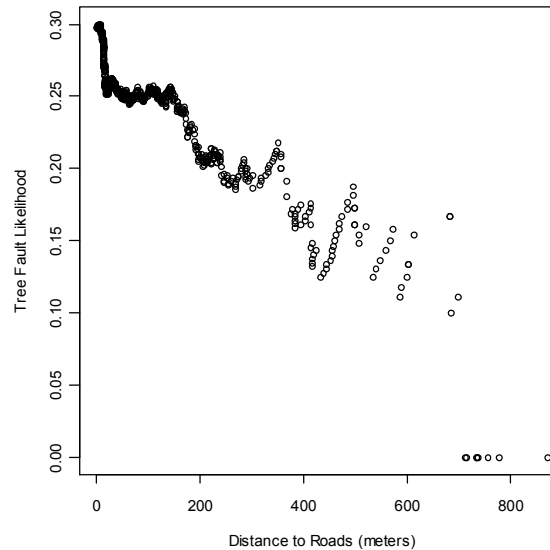
- Continuous features
  - Likelihood measure
  - plot

$$L_{i,j} = P(o_i | X \geq x_j) = \frac{N_{i,j}}{N_j}$$

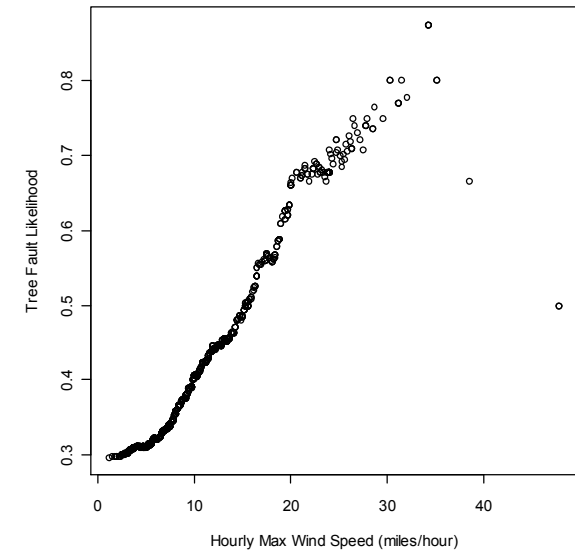
Distance to Trees vs. Tree Faults



Distance to Roads vs. Tree Faults



Wind Speed vs. Tree Faults



- Linear discriminant analysis (LDA)

$$D = \mathbf{w}^T \mathbf{f} = \sum_{i=1}^N w_i f_i \quad D = [\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)]^T \mathbf{f}$$

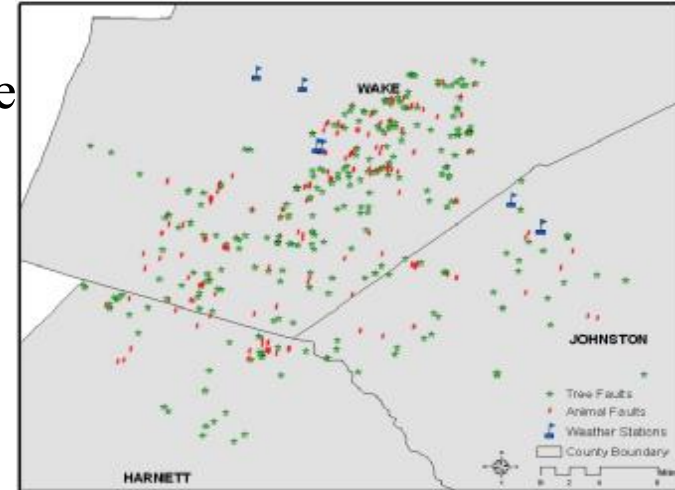
- Logistic regression (LR)

$$\text{logit}(c = 1) = \ln \frac{P(c = 1)}{P(c = 0)} = \alpha + \boldsymbol{\beta}^T \mathbf{f} \quad P(c = 1) = \frac{1}{1 + e^{-\alpha - \boldsymbol{\beta}^T \mathbf{f}}}$$

- Comparison

	LDA	LR
Model	linear classifier	non-linear classifier
Data assumption	normal distributed with equal variance	none
Computation	matrix manipulation	maximum likelihood

- Data sources
  - Progress Energy Carolinas outage database
  - NC Climate Office
  - NC State Univ. GIS data service



- Fault causes of interest
  - Tree-caused
  - Animal-caused
  - Other

- Features
  - 7 categorical
  - 5 continuous

- Classifiers
  - LDA
  - LR

Classification Performance Using LDA on Sample Dataset

		6 Features		12 Features	
		training	testing	training	testing
Tree fault	ACC	0.75(0.01)	0.76(0.01)	0.77(0.02)	0.76(0.02)
	POD	0.32(0.03)	0.34(0.03)	0.41(0.03)	0.39(0.03)
	FAR	0.34(0.03)	0.32(0.03)	0.32(0.04)	0.33(0.04)
Animal fault	ACC	0.84(0.02)	0.83(0.01)	0.84(0.02)	0.84(0.01)
	POD	0.31(0.04)	0.29(0.04)	0.35(0.03)	0.35(0.03)
	FAR	0.42(0.05)	0.43(0.05)	0.39(0.05)	0.41(0.05)

Classification Performance Using LR on Sample Dataset

		6 Features		12 Features	
		training	testing	training	testing
Tree fault	ACC	0.76(0.02)	0.76(0.02)	0.77(0.01)	0.77(0.02)
	POD	0.32(0.03)	0.32(0.03)	0.44(0.03)	0.44(0.03)
	FAR	0.30(0.04)	0.30(0.04)	0.32(0.03)	0.34(0.03)
Animal fault	ACC	0.83(0.02)	0.83(0.02)	0.84(0.01)	0.84(0.01)
	POD	0.30(0.03)	0.31(0.03)	0.37(0.04)	0.35(0.03)
	FAR	0.42(0.04)	0.41(0.06)	0.41(0.06)	0.41(0.06)

- Methods for exploratory data analysis
  - Integrate data from multiple sources under GIS framework
  - Use likelihood measure to evaluate both categorical and continuous features
  - Apply LDA and LR as fault cause classifiers
  
- Findings
  - LDA and LR performs similar
  - Adding new features helps fault diagnosis
  
- Future work
  - Systematic feature selection methods
  - Advanced fault diagnosis algorithms
  - Novel sampling strategy