



The Southern African Grain Laboratory NPC

Quality is our passion



DATA MINING OF ELEVEN YEARS' CROP QUALITY SURVEY RESULTS

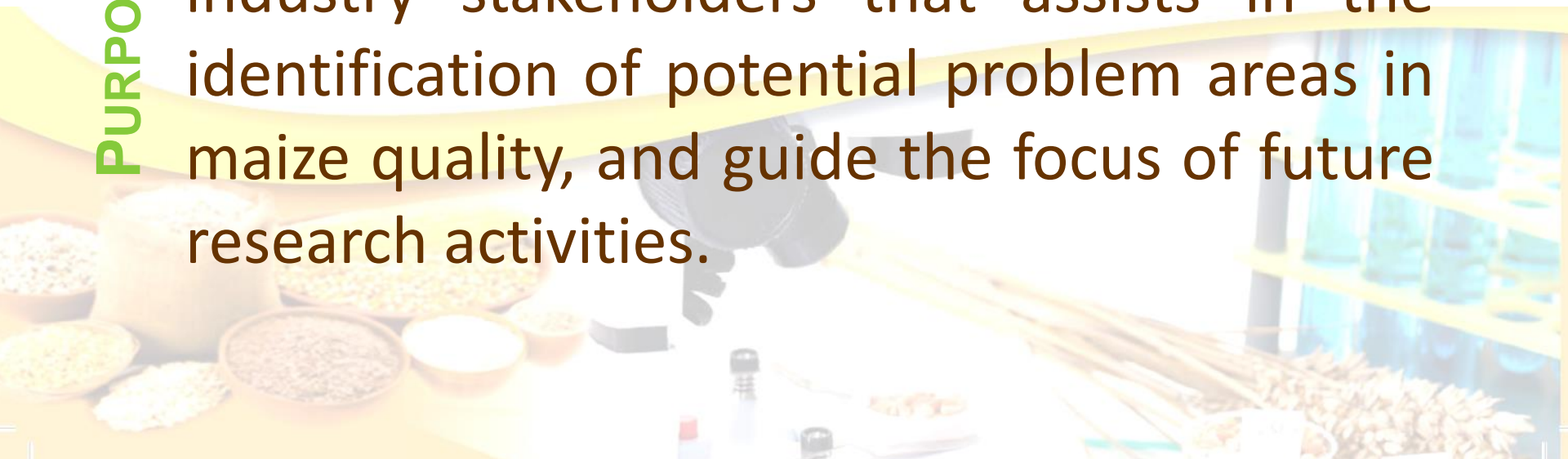
Dr. Corinda Erasmus

Ms. Wiana Louw



GOAL Evaluate crop quality data in order to identify unique South African trends.

PURPOSE To provide a decision making tool for maize industry stakeholders that assists in the identification of potential problem areas in maize quality, and guide the focus of future research activities.



DATASET CHALLENGES

- ✓ Crop quality data for eleven seasons are summarized in tables (White and Yellow maize separate)
- ✓ Datasets are skew (some regions > 300 sampling points; others < 50).
- ✓ A professional statistician (Wits/Monash University affiliated) recommended:
 - ✓ Create a sub-sample of the dataset by random sampling of the original set and then create a balanced workset.
- ✓ **Results are influenced by:**
 - ✓ **factors** - season, region and colour
 - ✓ **continuous traits** - % protein, starch, milling index, etc.

CHOOSING A STATISTICAL TEST

- ✓ Multifactor ANOVA - test the factors only
- ✓ Principal Component Analysis (PCA) or regression - preferred for the continuous datasets
- ✓ Classification and Regression Trees for a holistic view of all the effects (it combines ANOVA and Regression tests)
- ✓ Best practice to apply all for large incomplete datasets typically found in data mining applications as the tests often complement each other but none of them can give all the answers on their own
- ✓ **Know what question needs to be answered.**

CROP QUALITY DATASET

FREQUENCY TABLE OF WHITE MAIZE DATA, ELEVEN SEASONS

Summary Stub-and-Banner Table (White maize data with milling tests no blanks) Marked cells have counts > 10 (Marginal summaries are not)

Region	Season (2)	Season (3)	Season (4)	Season (5)	Season (6)	Season (7)	Season (8)	Season (9)	Season (10)	Season (11)	Season (12)	Row (Total)
10	2	7	3	3	11	1	0	0	0	0	0	27
11	5	14	9	1	8	9	0	1	2	0	1	50
12	18	16	13	16	14	16	5	15	11	21	13	158
13	9	9	8	11	34	13	8	17	9	29	41	188
14	29	35	21	27	46	39	18	36	18	31	47	347
15	21	15	9	13	13	33	2	25	19	10	13	173
16	26	22	13	23	13	15	19	41	22	14	30	238
17	11	23	23	26	31	21	17	33	14	30	27	256
18	16	40	26	19	30	17	19	10	14	21	33	245
19	14	16	15	16	13	18	10	10	10	25	25	172
20	12	9	15	10	20	9	13	17	10	13	11	139
21	25	32	29	43	8	24	19	24	29	9	31	273
22	31	37	47	53	16	33	16	19	10	10	40	312
23	33	35	139	66	46	98	47	30	34	27	51	606
24	26	29	70	41	30	36	45	16	32	21	22	368
25	24	18	14	12	19	18	17	12	23	18	6	181
26	15	17	19	17	29	6	14	19	19	10	7	172
27	12	9	12	13	4	7	1	6	6	2	2	74
28	18	20	20	23	31	17	11	25	40	25	17	247
29	16	14	12	27	10	15	39	31	33	0	19	216
30	22	22	3	26	32	21	48	22	18	7	32	253
31	8	11	1	15	0	7	5	12	14	2	8	83
32	11	12	5	22	3	21	24	19	26	16	27	186
33	12	14	0	11	28	20	29	1	4	6	24	149
34	25	18	51	43	42	25	28	12	16	31	32	323
35	15	8	7	8	8	10	14	9	8	5	4	96
36	8	10	8	14	7	7	15	17	16	15	13	130
Total	464	512	592	599	546	556	483	479	457	398	576	5662

RANDOM SUB-SAMPLING OF WHITE MAIZE, BALANCED DATASHEET

Summary Stub-and-Banner Table (Spreadsheet2)

Marked cells have counts > 10

(Marginal summaries are not marked)

Region	Season 2	Season 3	Season 4	Season 5	Season 6	Season 7	Season 8	Season 9	Season 10	Season 11	Season 12	Row Total
12	15	12	12	11	12	12	4	11	8	16	7	120
13	6	6	5	6	19	9	6	11	8	19	25	120
14	14	11	6	7	15	13	7	12	8	10	17	120
15	15	10	6	8	8	22	2	19	12	7	11	120
16	12	8	8	9	13	7	10	21	13	7	13	121
17	4	10	13	13	14	11	8	18	5	12	12	120
18	9	20	12	14	15	6	10	4	7	8	14	119
19	9	12	12	12	12	13	5	8	8	15	16	122
20	10	8	14	10	14	7	12	14	8	12	11	120
21	15	15	13	26	6	8	4	8	9	6	15	125
22	10	14	13	21	12	18	6	5	6	4	16	125
23	8	5	21	15	15	19	5	7	8	6	10	119
24	8	7	24	17	17	10	16	5	7	7	6	124
25	19	10	11	10	9	8	12	9	15	14	3	120
26	12	16	12	10	24	4	8	13	11	6	5	121
27	12	9	12	13	4	7	1	6	6	2	2	74
28	9	13	8	7	14	6	6	11	25	15	6	120
29	9	6	4	18	6	10	19	19	20	0	9	120
30	13	13	1	15	15	7	24	11	5	3	13	120
31	8	11	1	15	0	7	5	12	14	2	8	83
32	9	3	5	16	2	14	17	11	19	9	15	120
33	6	7	0	10	21	17	23	1	3	6	17	111
34	10	6	23	18	13	8	9	4	7	9	13	120
35	15	8	7	8	8	10	14	9	8	5	4	96
36	8	7	7	14	7	7	15	16	13	13	13	120
Total	265	247	250	323	295	260	248	265	253	213	281	2900

GIS SOFTWARE DEVELOPMENT FOR DATA MINING

- ✓ Historically - data presented in table format showing mean values and standard deviations of results for region – this is not ideal as it is difficult to see the big picture
- ✓ Explored the possibility of developing a GIS map system, where grain production regions are presented on a map of South Africa, with the boundaries illustrated
- ✓ The results of the crop quality traits can then be represented in a colour scale format - highest values the darkest colour and lowest values the lightest colour
- ✓ Mean values shown as a legend
- ✓ SIQ (with additional data from GSI on the regional boundary specifications) created a prototype software package based on an open source GIS package (QGIS)

STATISTICS FOR GIS MAPS

ANOVA TESTS AND HOMOGENOUS GROUPS

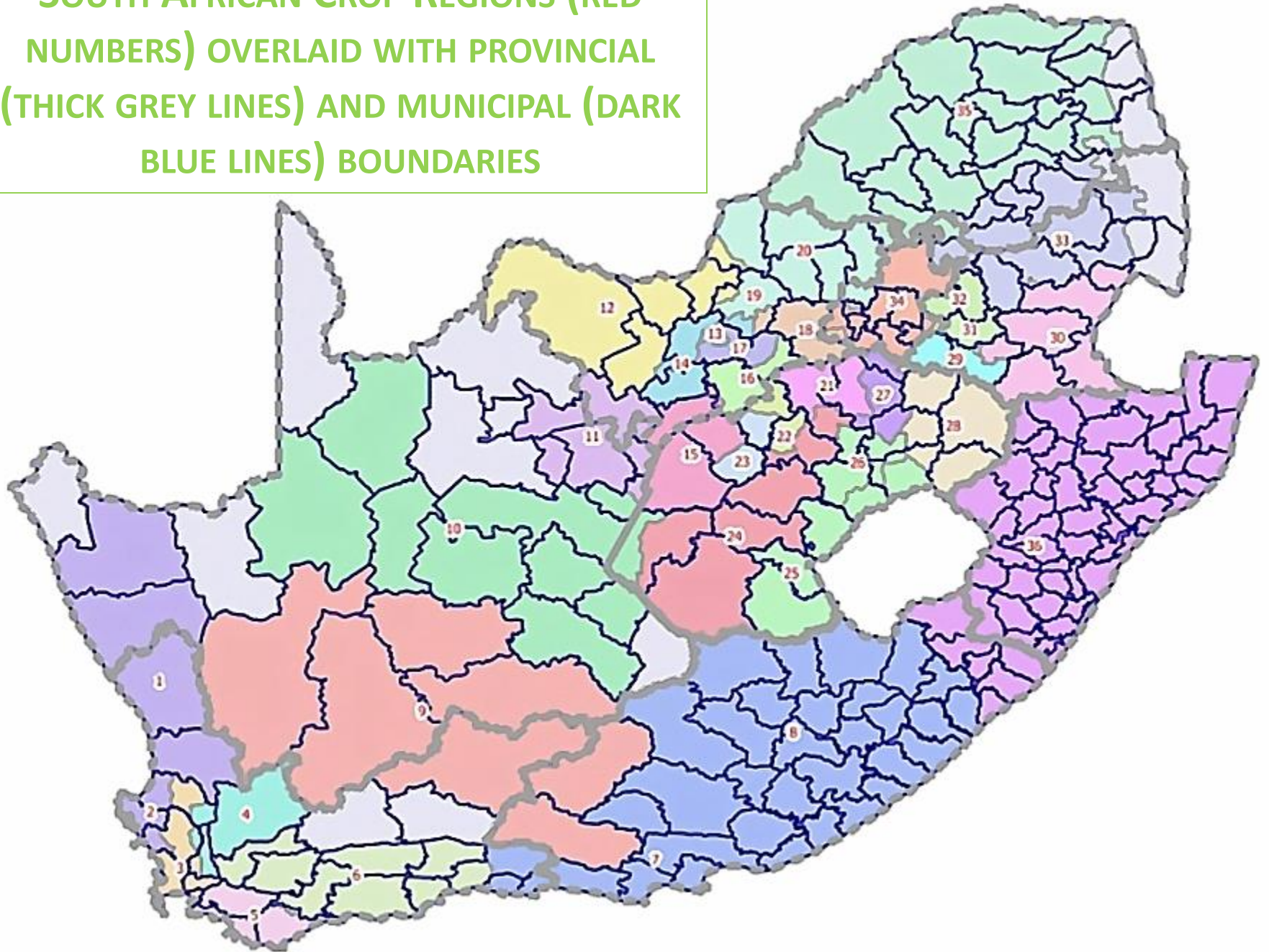
(POST HOC COMPARISONS) FOR MAPS

- ✓ Objective is to identify differences in samples
- ✓ Different types of ANOVA test – depending on the question asked.
- ✓ Looking for areas where specific traits are consistently higher or lower than the average or in comparison with other areas
 - ✓ For example, if a specific area always has the highest protein value irrespective of the season - points towards something influencing the value – for this we used a “liberal” ANOVA test
- ✓ Fisher LSD at 95% for the construction of the GIS maps
- ✓ GIS maps can show mean values for a trait for a specific region as:
 - ✓ Average for all seasons combined or
 - ✓ Individual seasons on a year to year basis

GIS MAPS

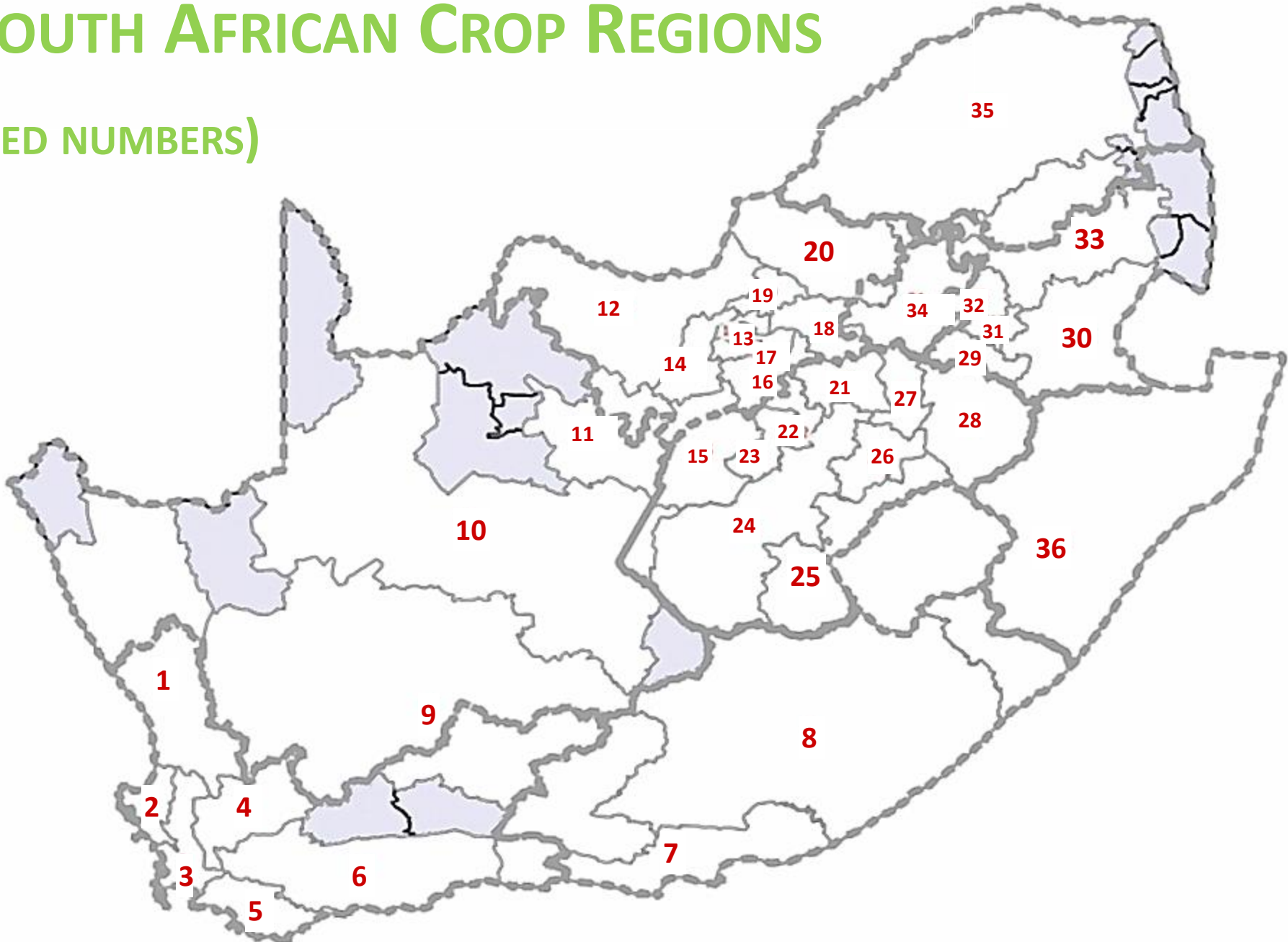
- ✓ Maps show mean values of the crop data per region
- ✓ Corresponding colour indicating the ranking of the means
- ✓ Areas with the same colour are **not statistically different**
- ✓ Light yellow areas are areas for which no data is available
- ✓ White areas are outside the Republic of South Africa
- ✓ Blue textured regions have no silo points
- ✓ Legends are indicated on each map
- ✓ Data at present available for 264 maps, but that is not all the grading data included and only goes to the 2011/2012 season. **Each season – adds 23 more maps!**

SOUTH AFRICAN CROP REGIONS (RED NUMBERS) OVERLAID WITH PROVINCIAL (THICK GREY LINES) AND MUNICIPAL (DARK BLUE LINES) BOUNDARIES



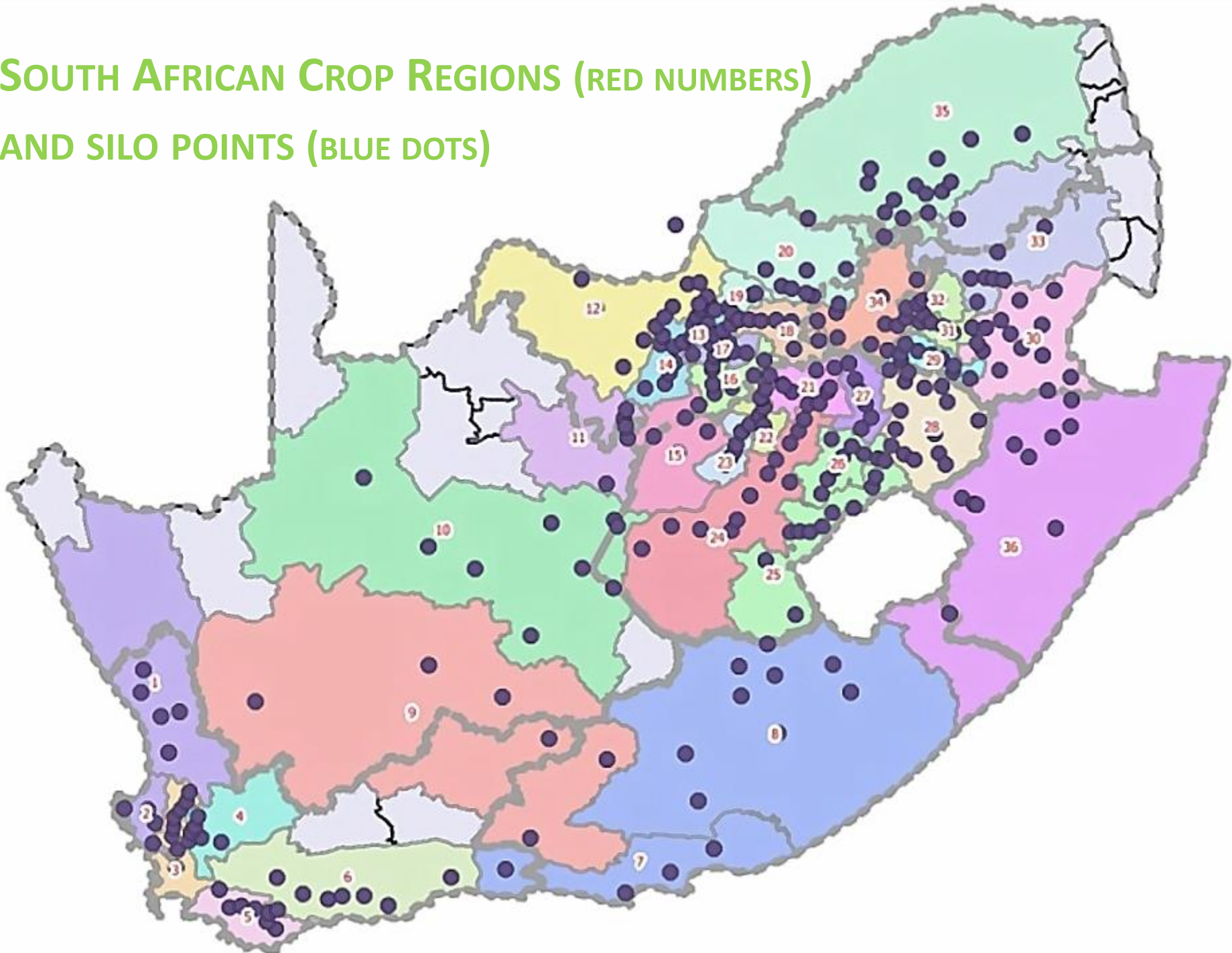
SOUTH AFRICAN CROP REGIONS

(RED NUMBERS)



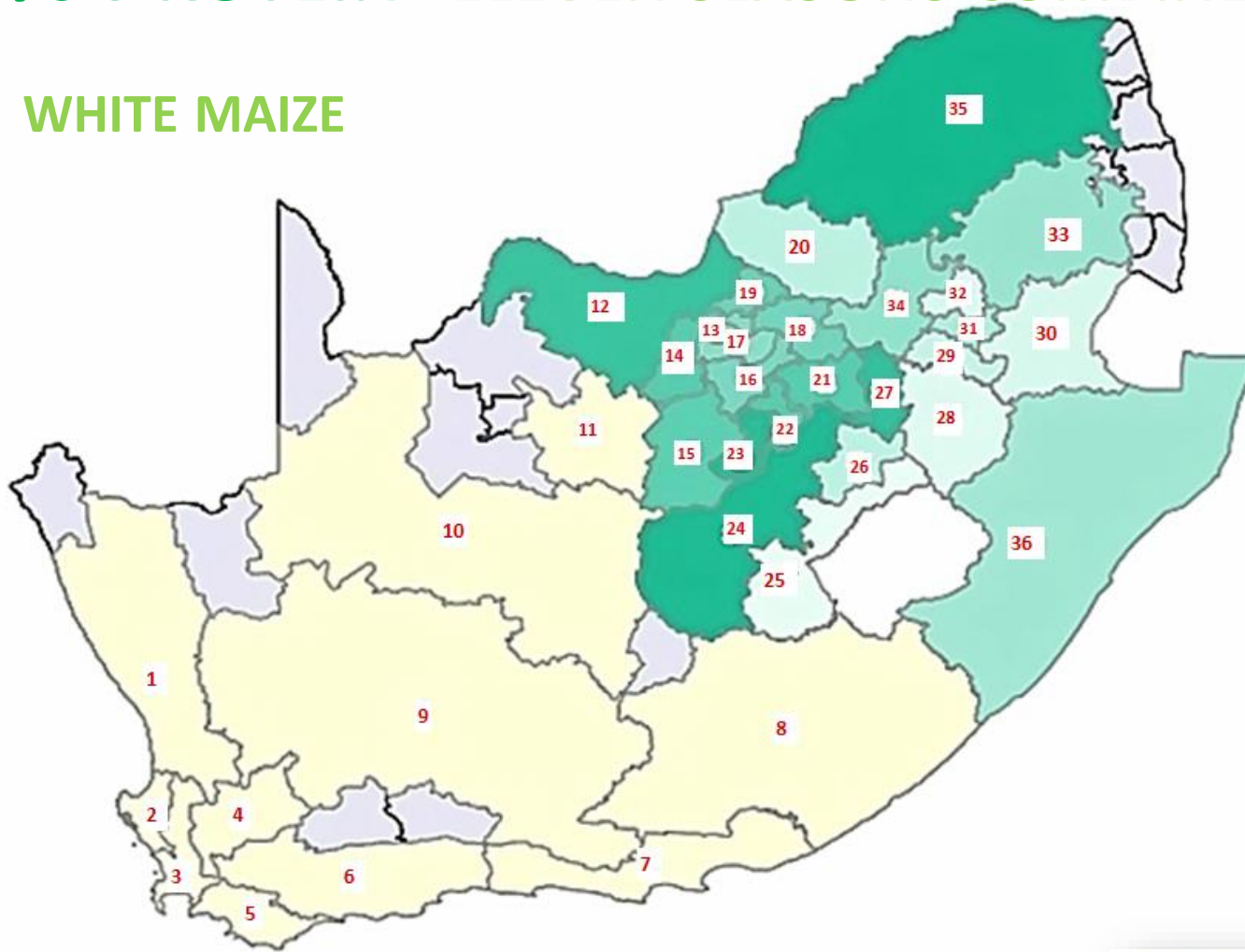
SOUTH AFRICAN CROP REGIONS (RED NUMBERS)

AND SILO POINTS (BLUE DOTS)



% PROTEIN ELEVEN SEASONS COMBINED

WHITE MAIZE



Legend

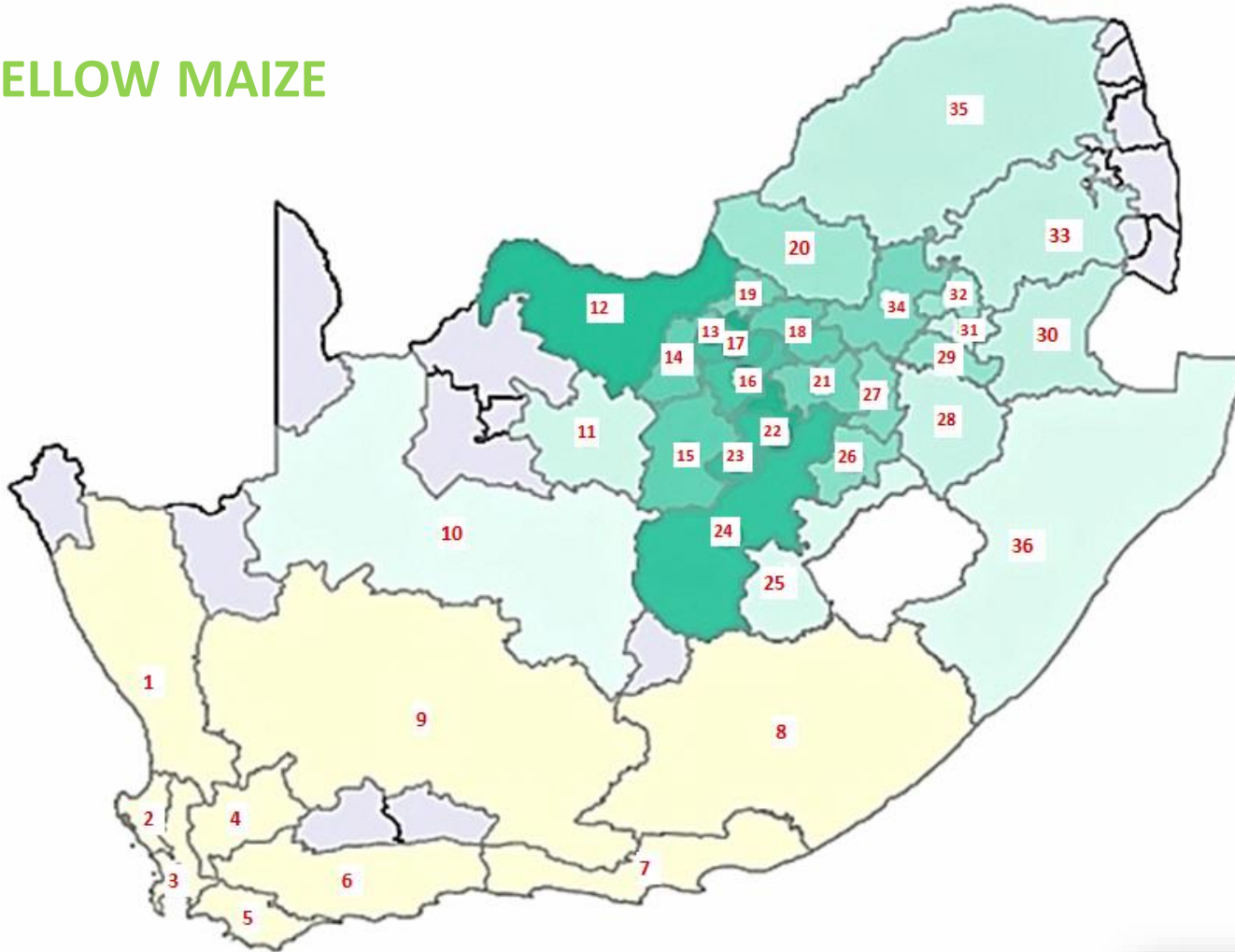
- Region_Label_Points
- RSA Crop Regions

LINK TO DB

- 0
- 8.19
- 8.31
- 8.43
- 8.44
- 8.48
- 8.48
- 8.49
- 8.53
- 8.53
- 8.56
- 8.64
- 8.69
- 8.71
- 8.71
- 8.73
- 8.82
- 8.87
- 8.87
- 8.88
- 8.88
- 8.90
- 8.92
- 9.01
- 9.08
- 9.08
- Municipal areas with no

% PROTEIN ELEVEN SEASONS COMBINED

YELLOW MAIZE



□ RSA Crop Regions

LINK TO DB

0

7.78

8.08

8.27

8.32

8.44

8.54

8.54

8.56

8.57

8.62

8.63

8.68

8.72

8.78

8.83

8.90

8.90

8.92

8.94

8.97

8.99

9.01

9.02

9.05

9.12

9.21

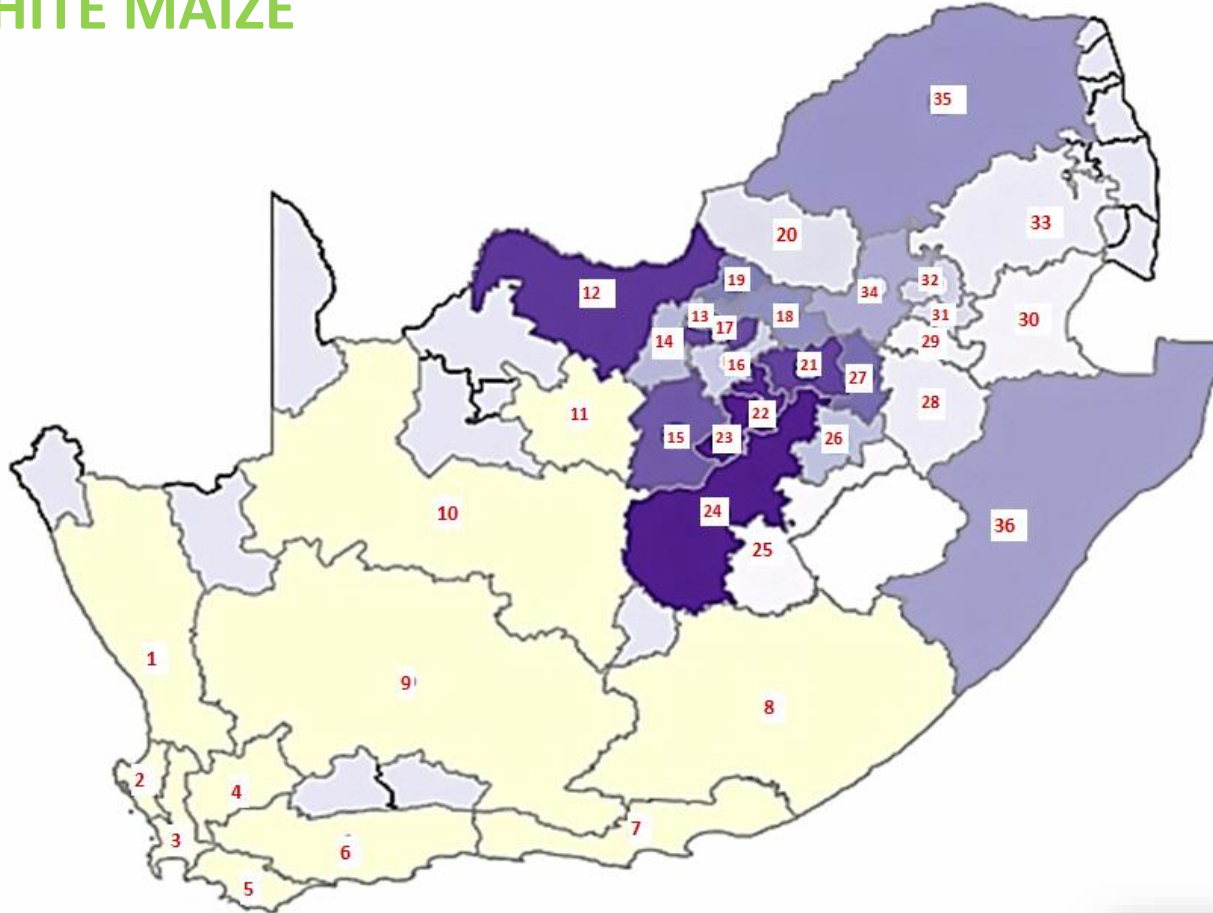
9.21

9.21

□ Municipal areas with no

ROFF MILLING % GRITS

WHITE MAIZE



Legend

Region_Label_Points

RSA Crop Regions

LINK TO DB

0

26.60

27.15

27.34

27.67

27.75

28.05

28.12

28.21

28.33

28.39

28.60

28.80

28.85

29.04

29.08

29.11

29.14

29.22

29.26

29.32

29.40

29.66

30.09

30.57

30.62

Municipal areas with no s

% STARCH ELEVEN SEASONS COMBINED

WHITE MAIZE

Legend

Region_Label_Points

RSA Crop Regions

LINK TO DB

0

72.84

72.91

73.12

73.16

73.23

73.23

73.27

73.29

73.29

73.30

73.30

73.39

73.40

73.45

73.49

73.51

73.52

73.54

73.58

73.69

73.78

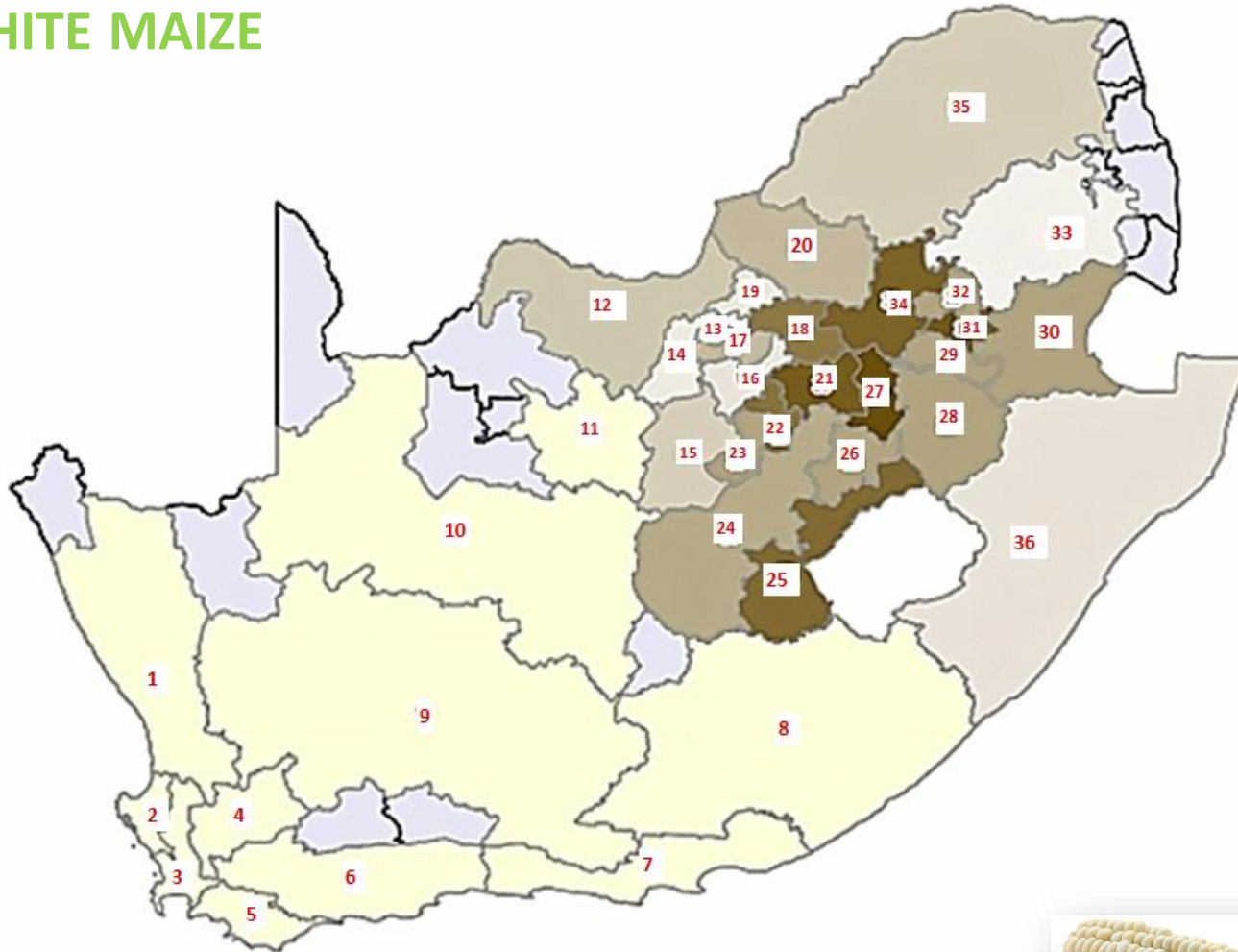
73.82

73.98

74.03

74.32

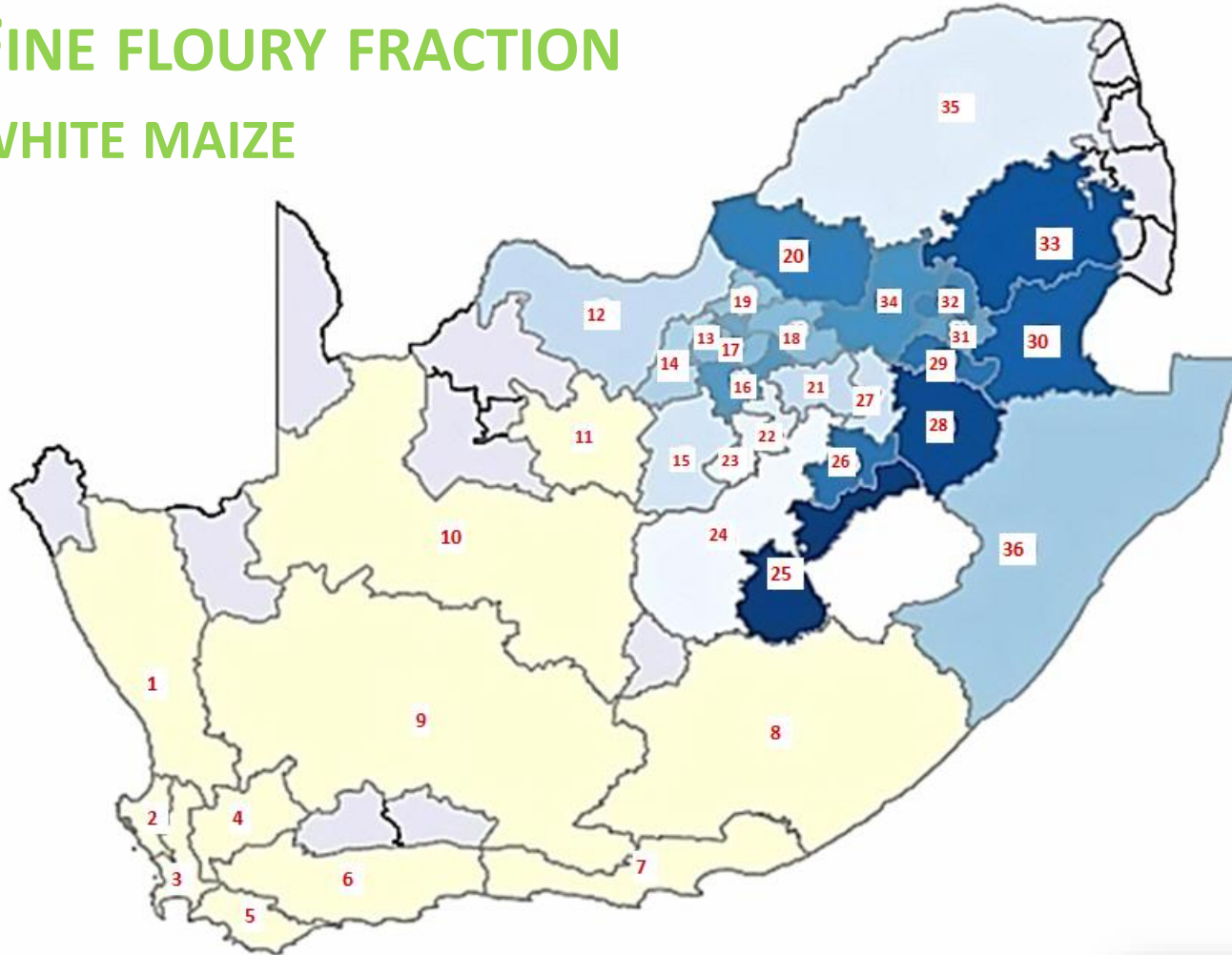
Municipal areas with no s



ROFF MILLING % BREAK 1

FINE FLOURY FRACTION

WHITE MAIZE



Legend

Region_Label_Points

□ RSA Crop Regions

LINK TO DB

□ 0

□ 12.15

□ 12.25

□ 12.43

□ 12.47

□ 12.59

□ 12.65

□ 12.66

□ 12.66

□ 12.78

□ 12.83

□ 12.85

□ 12.99

□ 13.03

□ 13.10

□ 13.19

□ 13.19

□ 13.34

□ 13.52

□ 13.59

□ 13.64

□ 13.73

□ 13.77

□ 13.86

□ 14.19

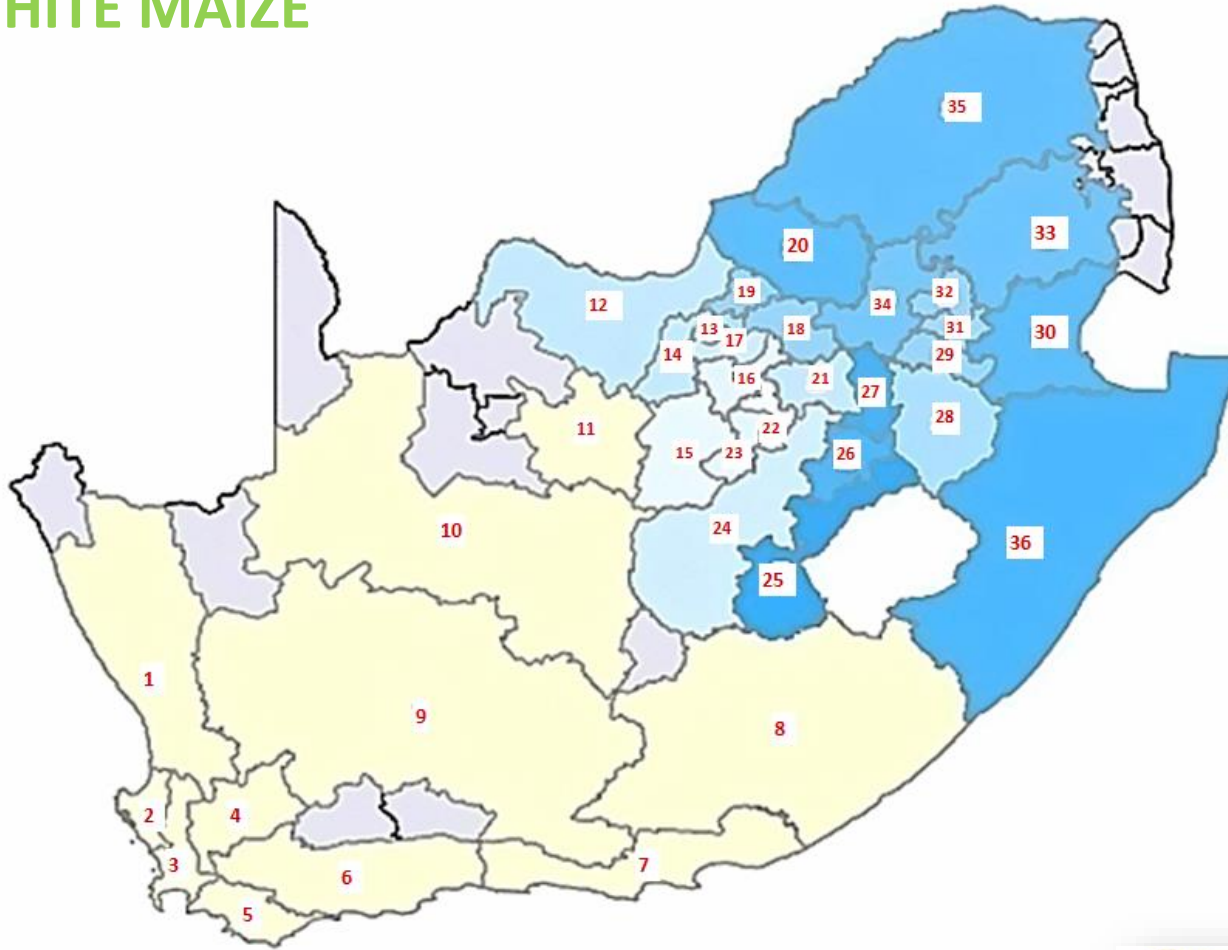
□ 14.63

□ Municipal areas with no



ROFF MILLING % BRAN

WHITE MAIZE



Legend

Region_Lable_Points

□ RSA Crop Regions

LINK TO DB

0

20.74

20.83

20.93

20.93

21.04

21.32

21.40

21.41

21.43

21.46

21.47

21.56

21.63

21.69

21.75

21.78

21.79

21.89

21.96

21.98

22.09

22.29

22.35

22.35

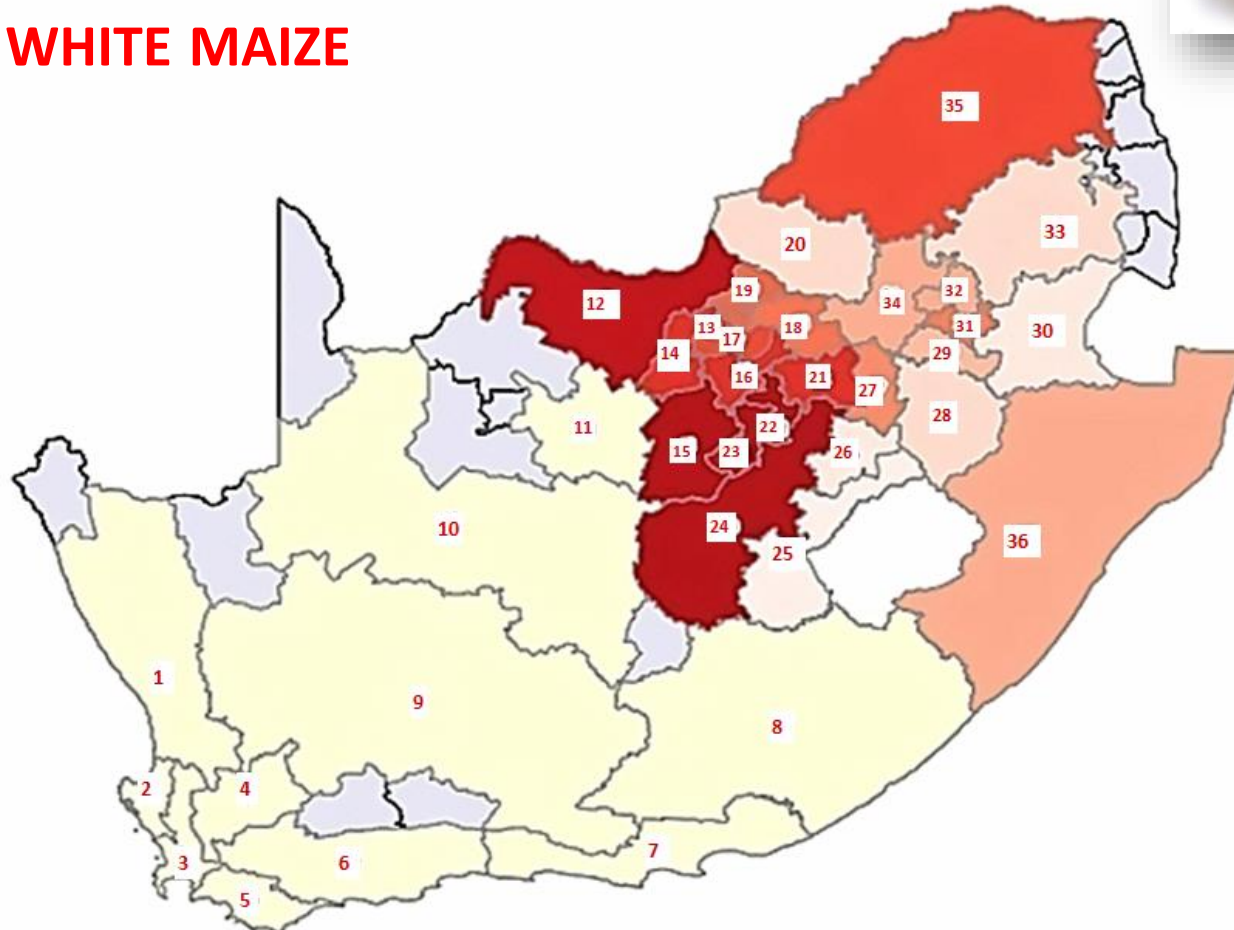
22.62

□

□ Municipal areas with no s

ROFF MILLING INDEX

WHITE MAIZE



Legend

Region_Label_Points
 RSA Crop Regions

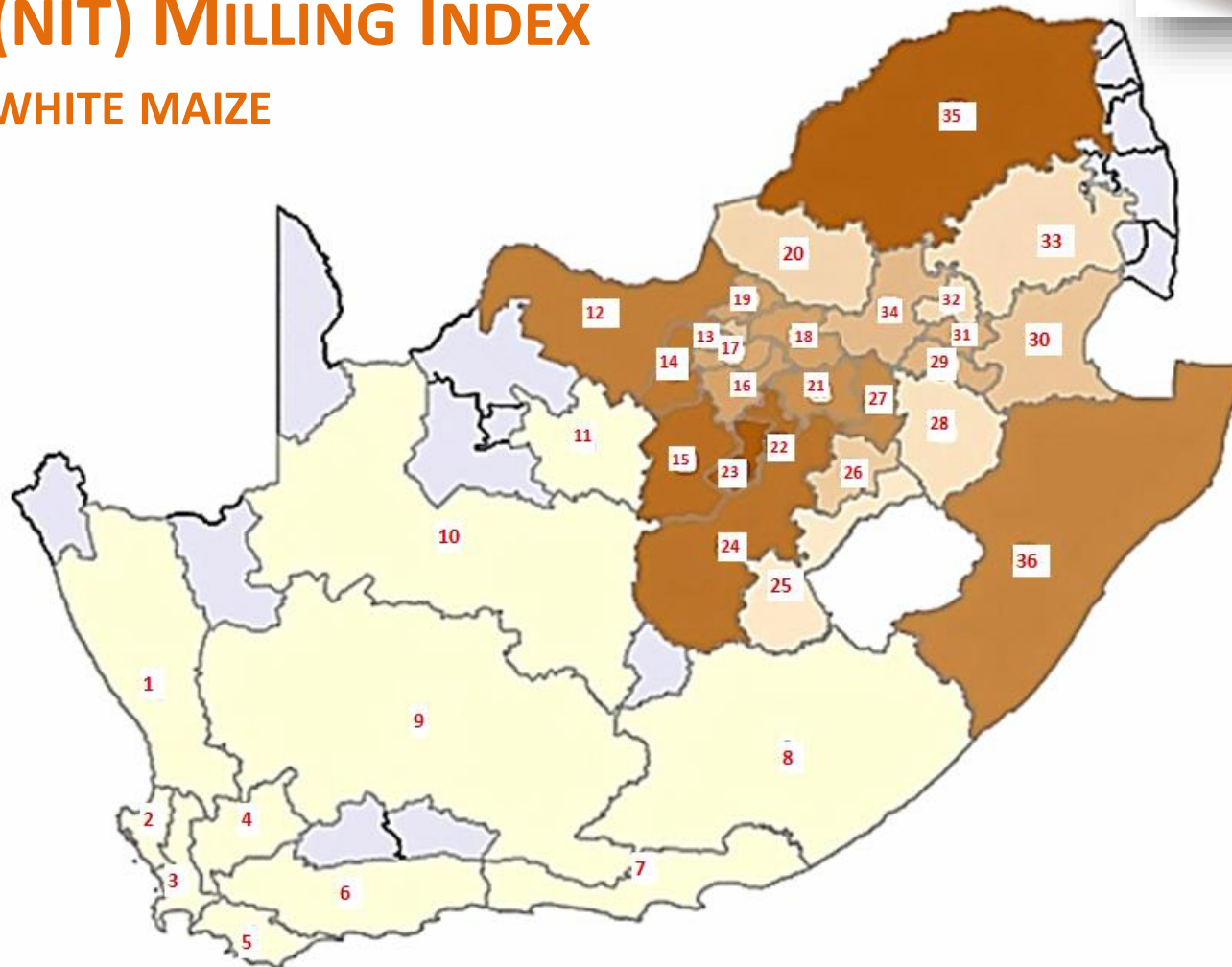
LINK TO DB

0
 71.08
 78.11
 78.84
 79.49
 80.00
 80.43
 81.67
 82.44
 82.91
 82.98
 83.62
 85.05
 85.10
 85.71
 86.00
 86.90
 87.08
 88.07
 88.56
 88.59
 88.84
 92.47
 93.62
 93.75
 93.77
 Municipal areas with no si

A HIGHER MILLING INDEX = HIGHER YIELD OF HIGH VALUE PRODUCTS FROM THE HARD ENDOSPERM

NEAR INFRARED TRANSMISSION (NIT) MILLING INDEX

WHITE MAIZE



Legend

Region_Label_Points

□ RSA Crop Regions

LINK TO DB

0

89.81

92.56

93.35

94.32

94.32

94.71

95.08

95.88

96.58

96.87

96.95

97.01

98.00

98.00

98.06

98.25

98.50

98.73

99.05

99.41

99.94

100.2

100.3

100.6

101.0

□

□ Municipal areas with no s

THE NIT IS DONE ON WHOLE MAIZE AND THE MILLING INDEX CALIBRATION WAS DONE TO PROVIDE A RAPID MILLING INDEX TEST FOR MAIZE INTAKE QUALITY CONTROL PURPOSES

GIS MAPS OF GRADING RESULTS

- ✓ This map shows certain regions with consistently high levels of defective kernels.
- ✓ It is possible to create similar maps of the types of defective or deviant kernels for example frost damage, water damage, fungal damage, broken kernels, discoloured kernels etc.
- ✓ Data in industry is not always kept at that level of detail because the grading regulations group the defects together which is followed by the seed graders.
- ✓ Regions with high milling index also had higher levels of defective kernels – not sure why.
- ✓ Not all defective kernels are necessarily diseased, some are only broken.

Legend

Region_Label_Points

RSA Crop Regions

LINK TO DB

0

3.96

4.67

4.73

4.73

4.78

4.87

5.06

5.16

5.16

5.20

5.26

5.28

5.29

5.46

5.57

5.57

5.68

5.68

5.69

6.02

6.25

6.39

6.63

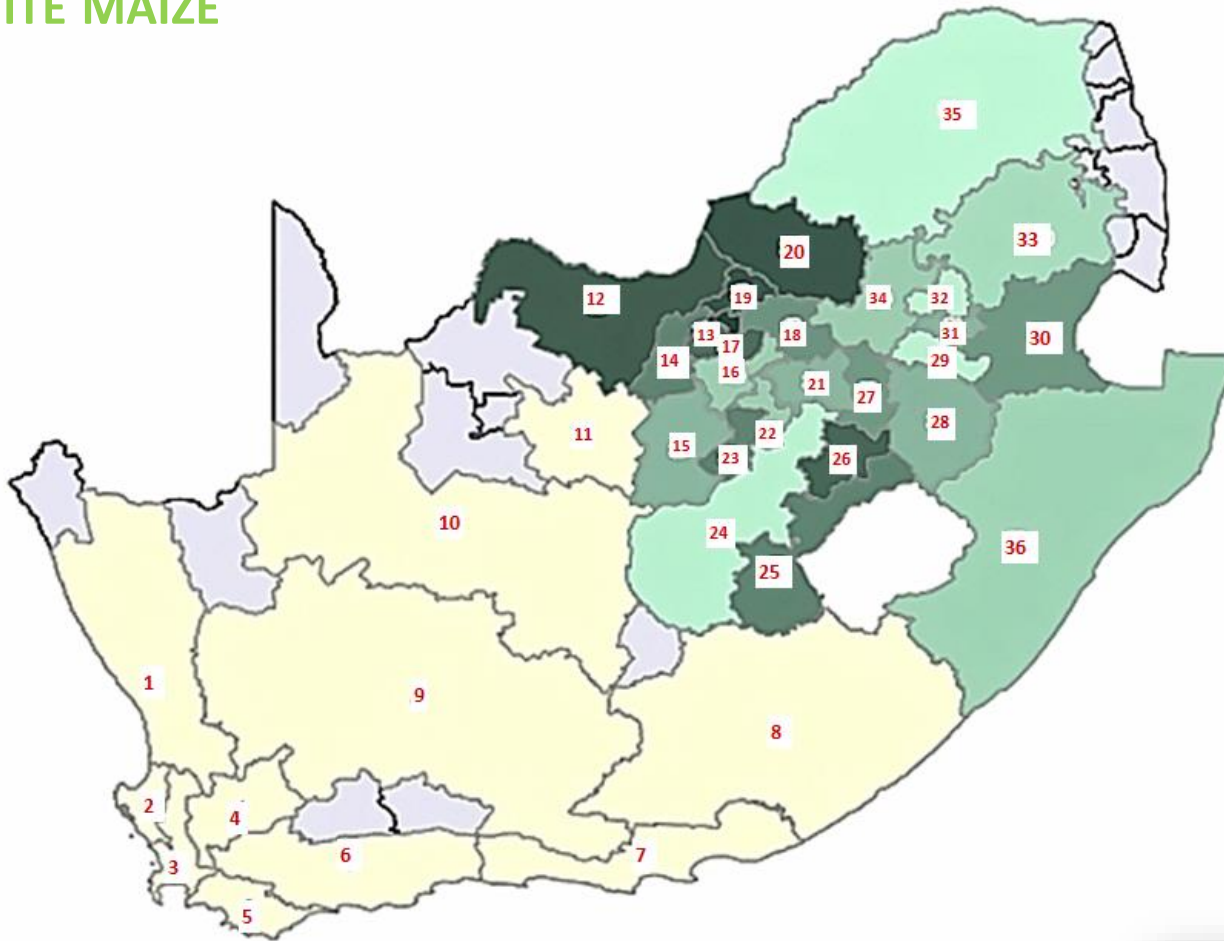
6.76

6.88

Municipal areas with no s

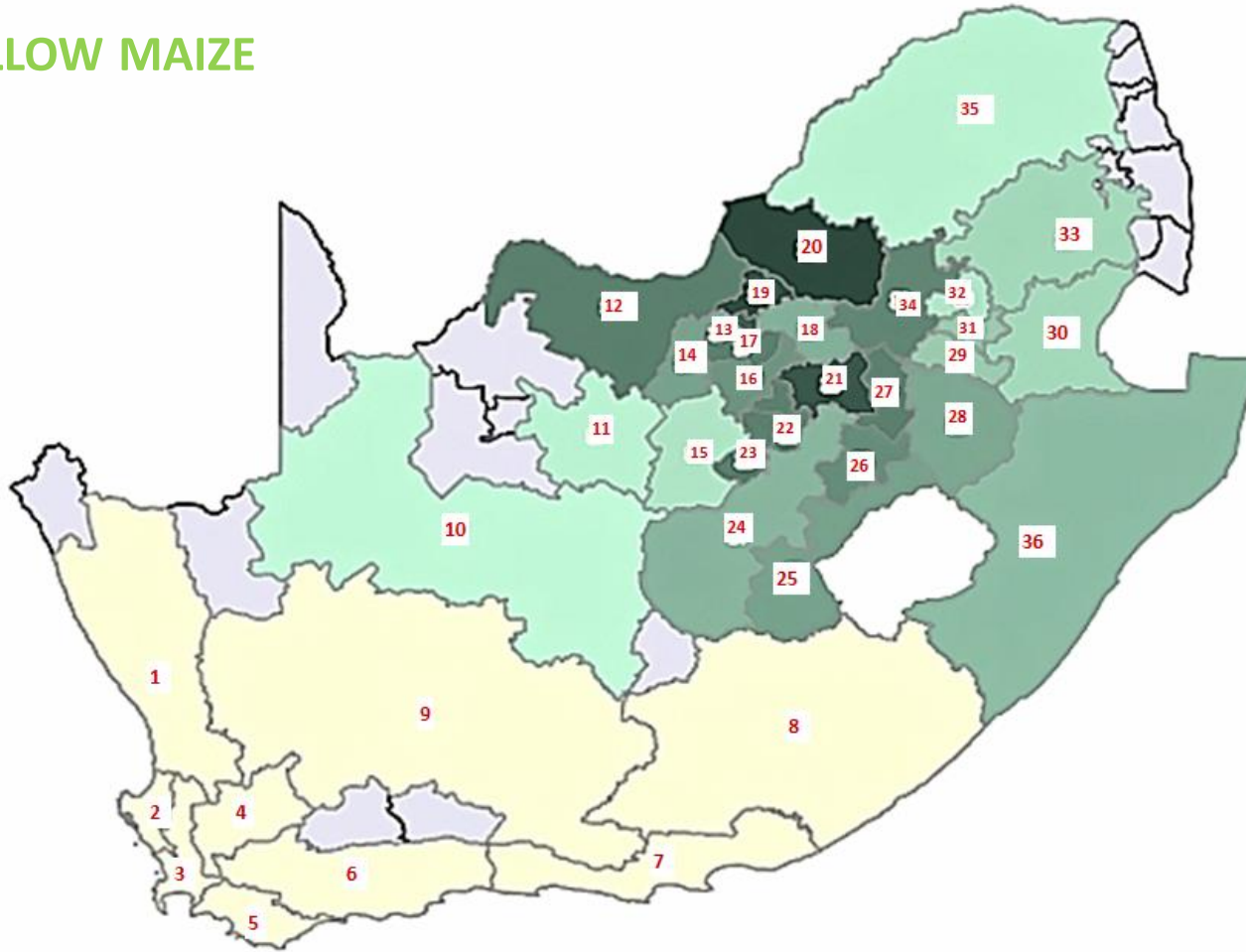
TOTAL DEFECTIVE ELEVEN SEASONS COMBINED

WHITE MAIZE



TOTAL DEFECTIVE ELEVEN SEASONS COMBINED

YELLOW MAIZE



Region_Label_Points

LINK TO DB

0

3.51

4.16

4.28

4.80

4.96

5.20

5.21

5.23

5.25

5.26

6.01

6.05

6.08

6.11

6.11

6.12

6.17

6.38

6.47

6.54

6.77

6.93

6.97

7.01

7.12

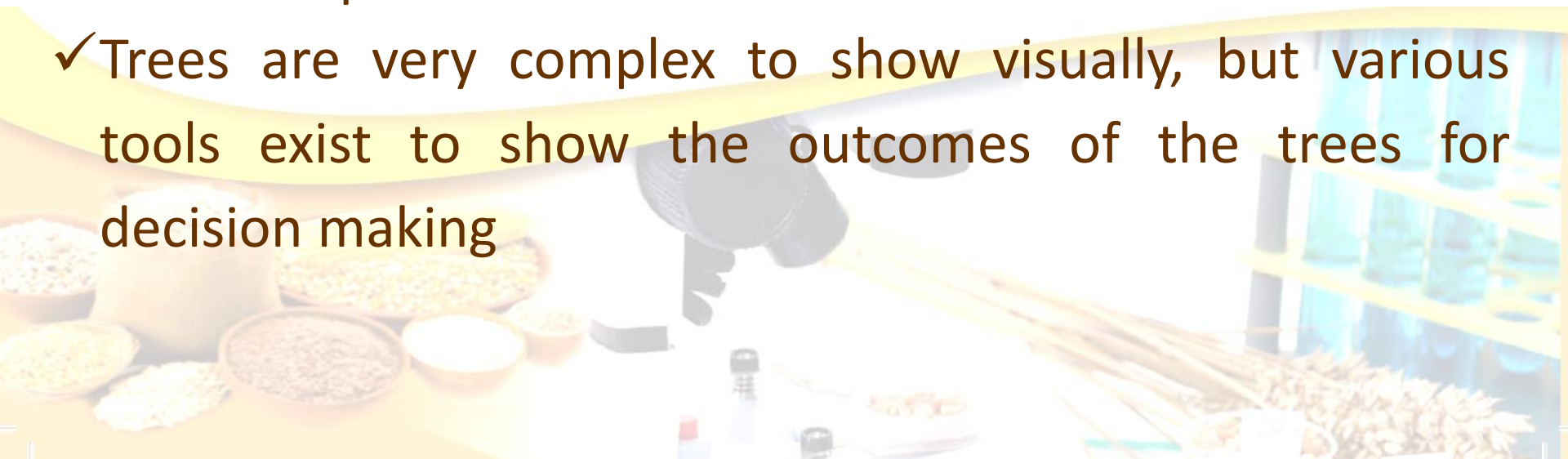
7.26

Municipal areas with no s



CLASSIFICATION AND REGRESSION TREES

- ✓ The tree example shows the effects of season, region, %protein, %starch, %fat, hectolitre mass, 100 kernel mass and total deviation on the yield of % Grits in white maize samples taken over eleven seasons (from 2001/2002 year to 2011/2012 year)
- ✓ 2900 samples were tested
- ✓ Trees are very complex to show visually, but various tools exist to show the outcomes of the trees for decision making

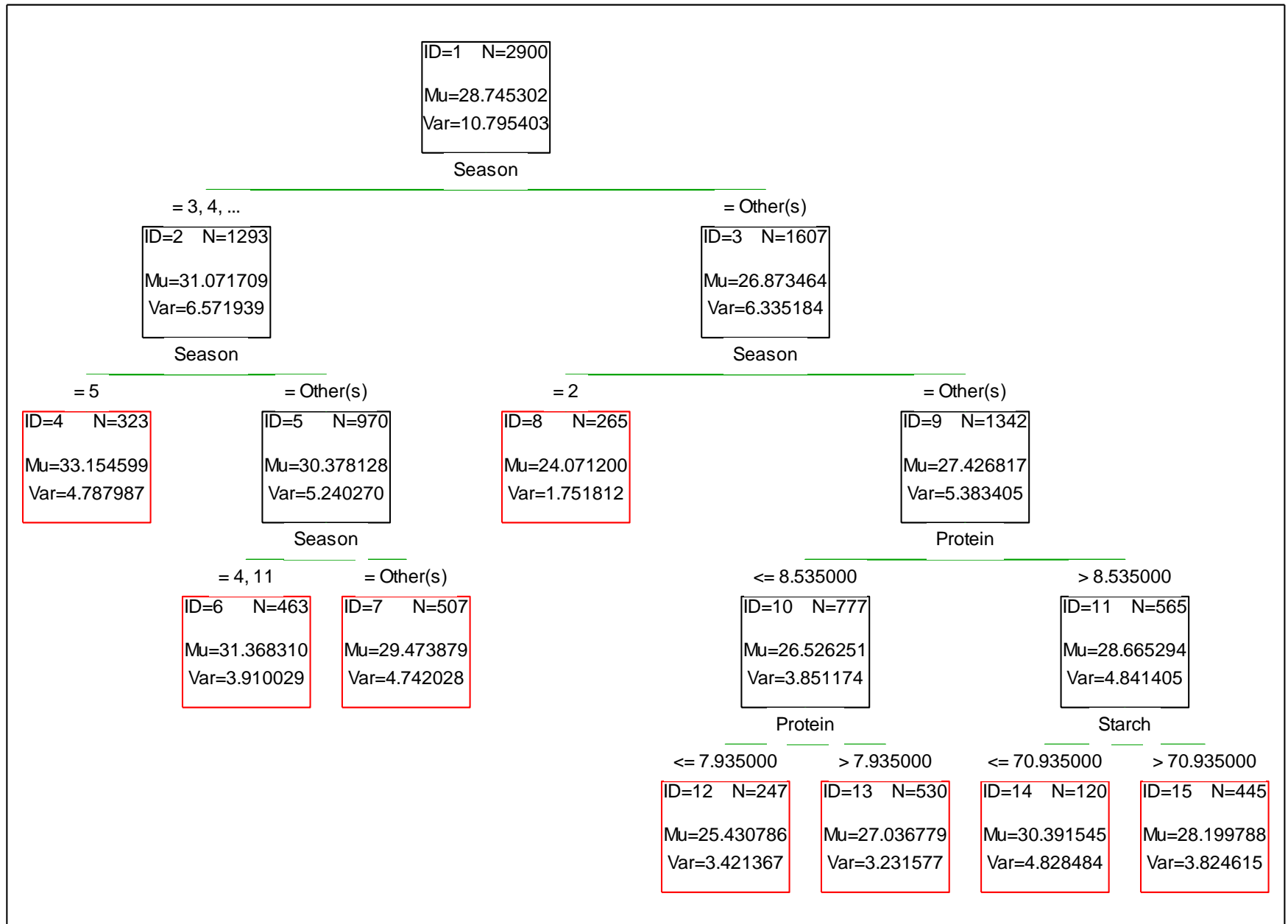


INDEPENDENT VARIABLE (PREDICTOR) IMPORTANCE PLOTS

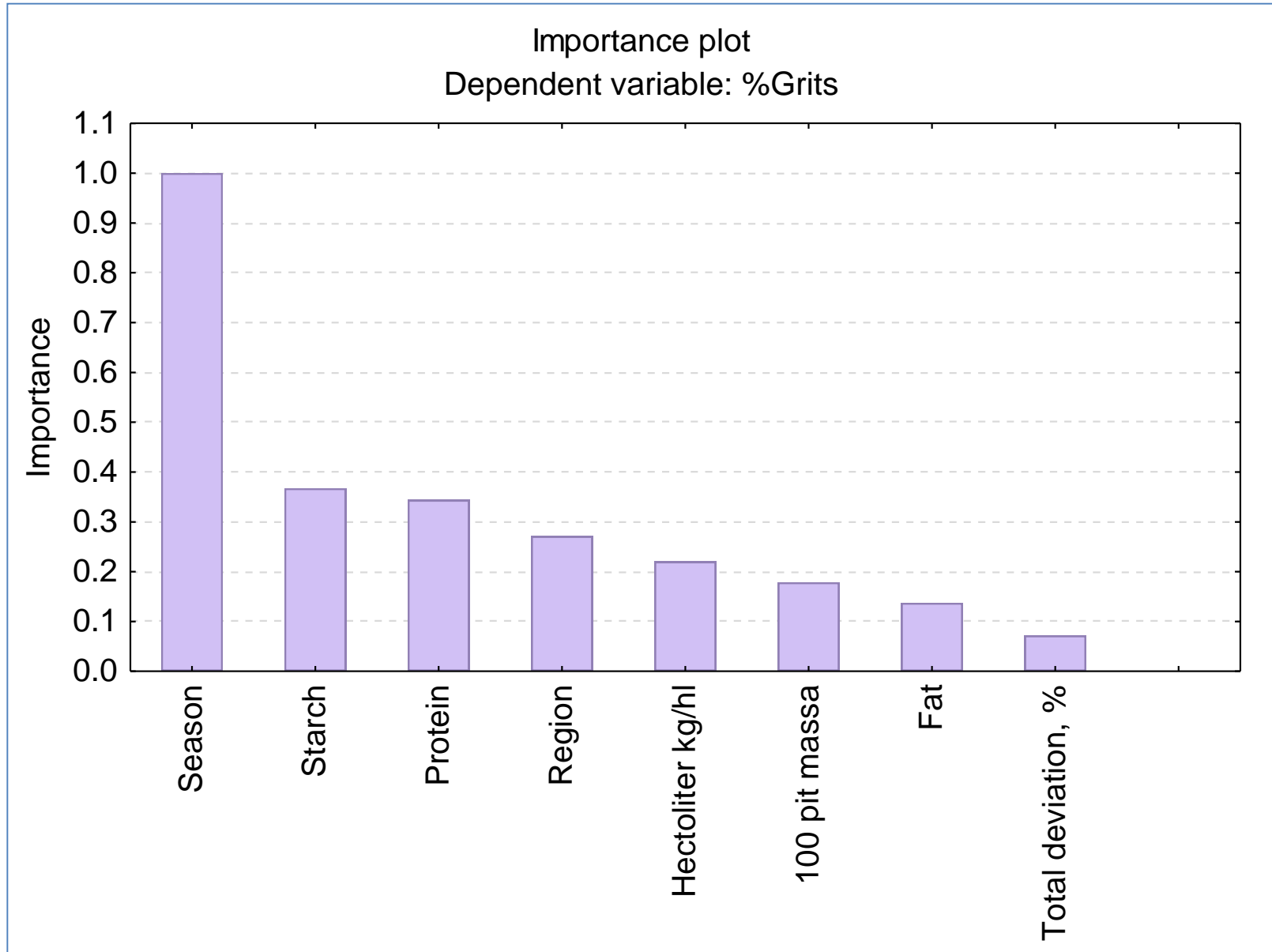
- ✓ Very useful tool to show quick result of a regression tree
- ✓ Shows which one of either the factors or the continuous variables are the main predictors of a specific trait
- ✓ Shows the factors and variables' relative importance to each other.

Tree 1 graph for %Grits

Num. of non-terminal nodes: 7, Num. of terminal nodes: 8

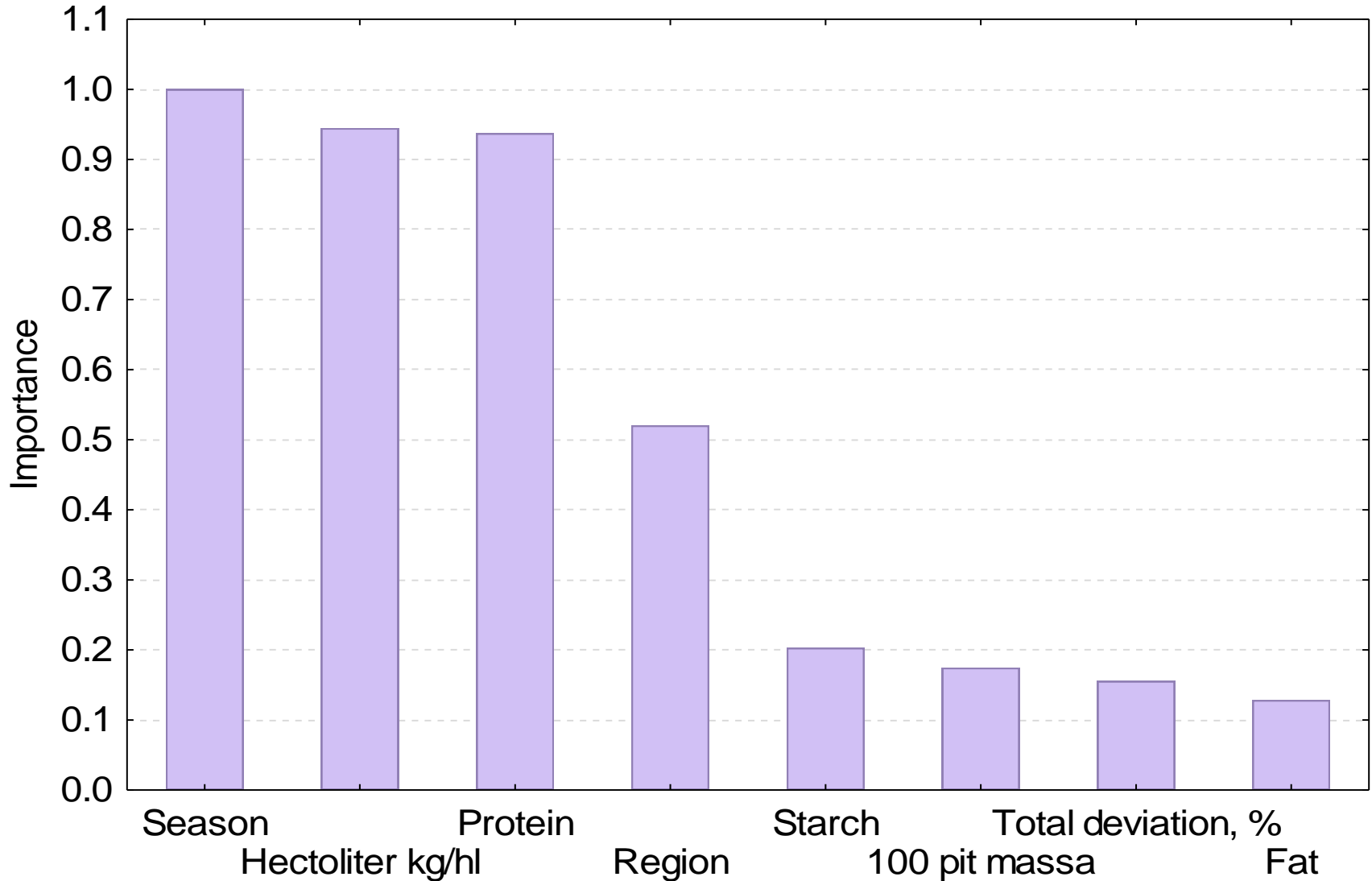


REGRESSION TREE IMPORTANCE PLOTS FOR % GRITS (MASS), WHITE MAIZE ONLY, N = 2900, ELEVEN SEASONS

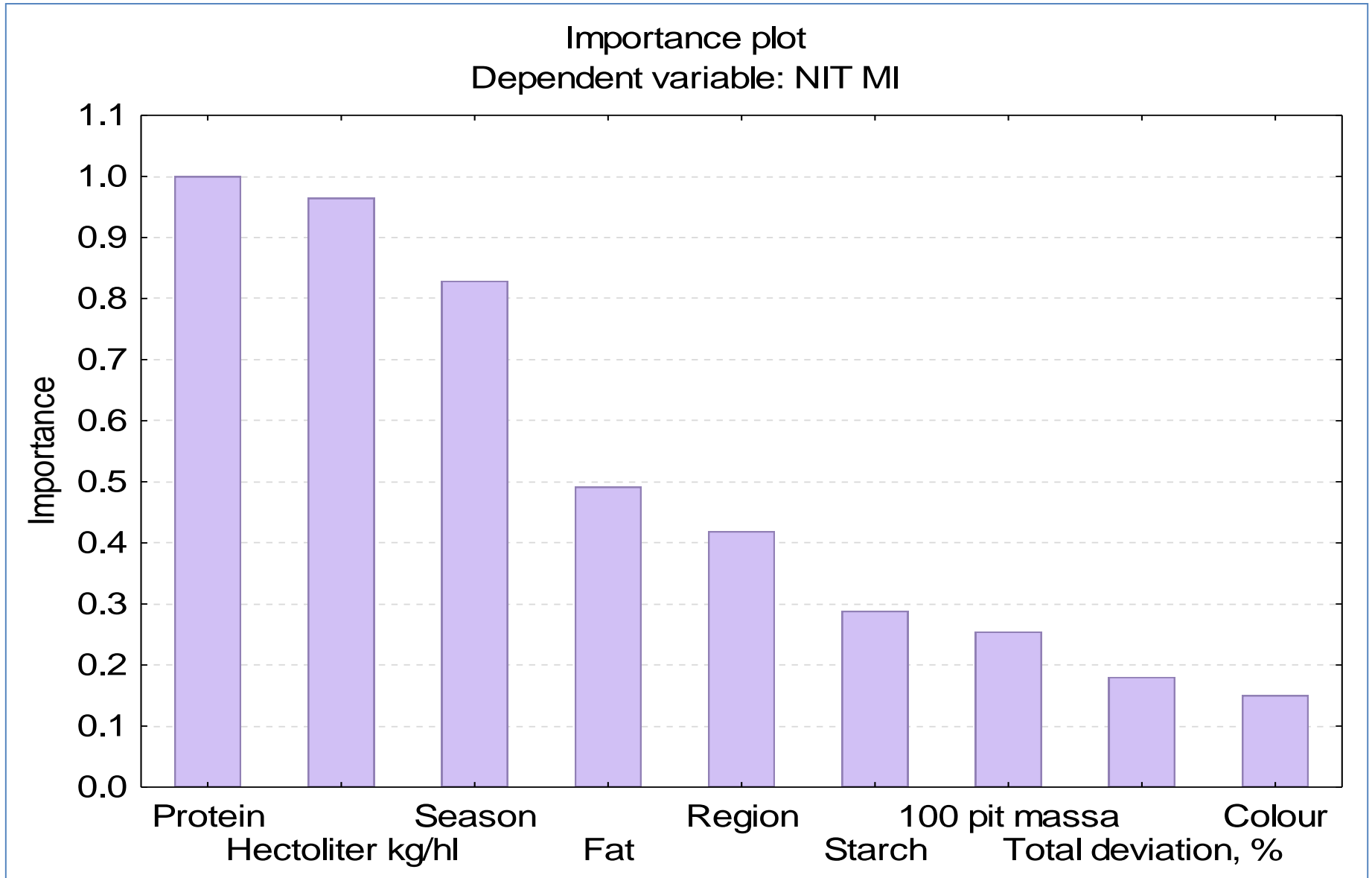


REGRESSION TREE IMPORTANCE PLOTS FOR ROFF MILLING INDEX, WHITE MAIZE ONLY, N = 2900, ELEVEN SEASONS

Importance plot
Dependent variable: NMI

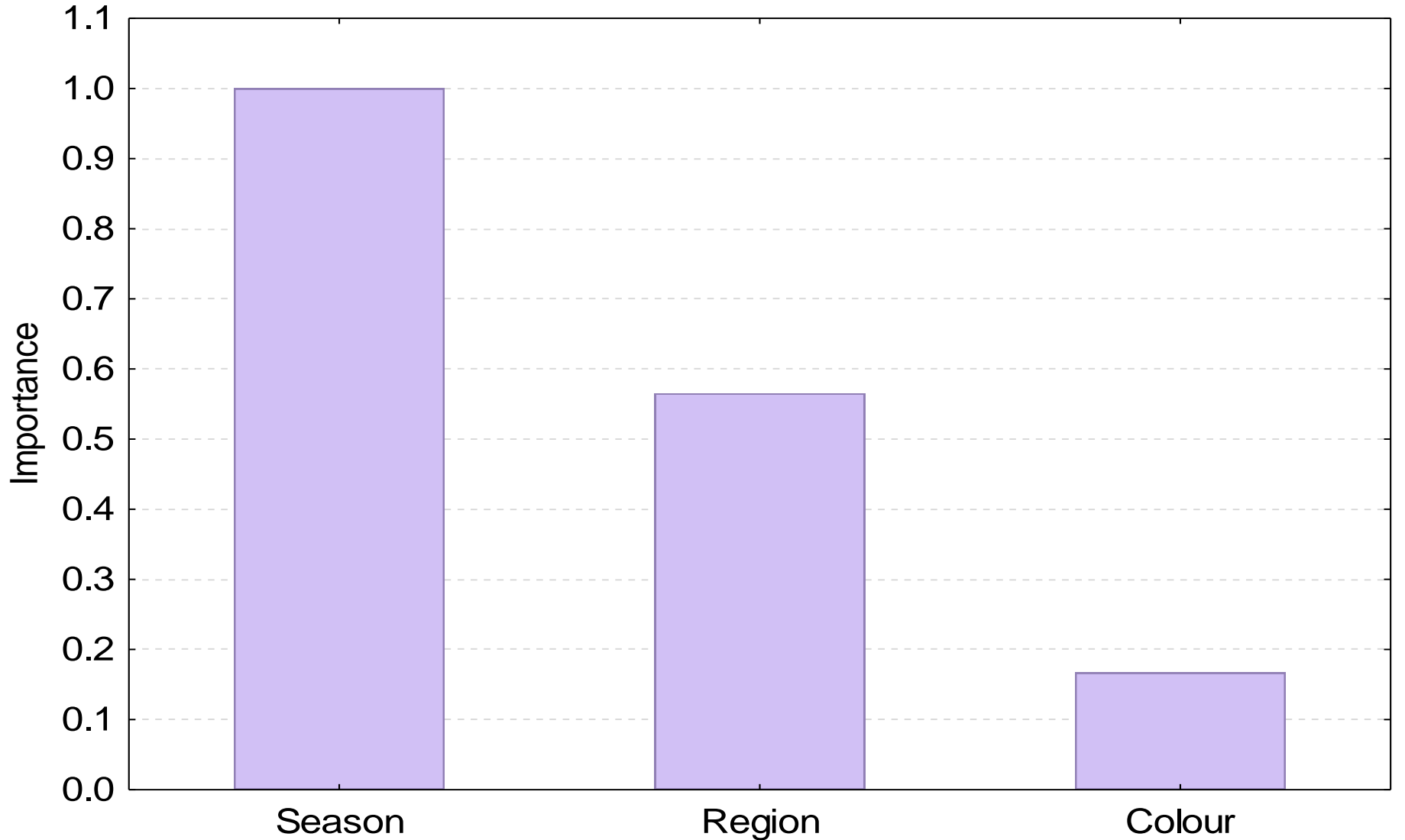


REGRESSION TREE IMPORTANCE PLOTS FOR THE CURRENT NIT MILLING INDEX CALIBRATION, WHITE MAIZE ONLY, N = 2900, ELEVEN SEASONS



All maize, eleven seasons example

Importance plot
Dependent variable: Total defective, %



CONCLUSION

- ✓ Data Mining provided a useful means to interpret the crop quality analytical data in a holistic way, by using a combination of statistical tools
- ✓ Identified trends to assist with future direction of decisions – many detailed discussions can follow from this work
- ✓ The goal of the project to present the data in a more accessible fashion has been achieved
- ✓ GIS tool – needs some further optimisation but it works very well and it has showed interesting trends that could not have been seen otherwise
- ✓ GIS tool can be expanded to other crops.

ACKNOWLEDGEMENTS

- ✓ The Maize Trust for funding of the research
- ✓ SAGL for diligent keeping of the many years' data in accessible electronic form
- ✓ Ms. Merle Werbeloff from Monash University for her assistance with the STATISTICA software usage and data mining techniques
- ✓ SIQ for their excellent work on the GIS map software
- ✓ The Silo Industry for providing the boundary information to SIQ in order to create these unique maps