The Southern African Grain Laboratory NPC
Quality is our passion

DATA MINING OF TWELVE YEARS' WHEAT CROP QUALITY SURVEY RESULTS

Dr. Corinda Erasmus
Ms. Jolanda Nortjé
Ms. Wiana Louw

SAGL
Southern African Grain Laboratory
SINCE 1997

GOAL

EVALUATE SEVERAL YEARS OF WHEAT CROP QUALITY DATA IN ORDER TO IDENTIFY UNIQUE SOUTH AFRICAN TRENDS.

PURPOSE

TO PROVIDE A DECISION MAKING TOOL TO THE WHEAT INDUSTRY STAKEHOLDERS ASSISTING IN THE IDENTIFICATION OF POTENTIAL PROBLEM AREAS IN WHEAT QUALITY AND THE FOCUS OF FUTURE RESEARCH ACTIVITIES.

MATERIALS AND METHODS

SAGL
Southern African Grain Laboratory
SINCE 1997

- ✓ Quality data for twelve seasons were analysed using the software and methods developed previously for maize crop quality data mining
- ✓ Two sets of models were developed namely for KOK (Koring Oes Kwaliteit) and SKOK (Saamgestelde Koring Oes Kwaliteit)
- ✓ Sub-samples of the datasets were used for statistics - random sampling were done from the original set to create a balanced worksheet
- ✓ Results are influenced by
  - ✓ **factors** - season, region
  - ✓ **continuous traits** - %protein, starch, rheology parameters, % deviation (grading) etc.

DATA MANIPULATION OPTIONS

- ✓ Multifactor **ANOVA** - test the factors only
- ✓ Principal Component Analysis (**PCA**) or regression - preferred for the continuous datasets (both traditional regression techniques or Partial Least Squares (**PLS**) Regression – modern)
- ✓ Classification and Regression Trees (**C&RT**) 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 to identify repeatable trends

GIS SOFTWARE DEVELOPMENT FOR DATA MINING

SAGL
Southern African Grain Laboratory
SINCE 1997

- ✓ Data historically presented in table format - this is difficult to interpret
- ✓ GIS map system was successfully developed for the maize crop quality data a few years ago
- ✓ The results of the crop quality traits were represented in a colour scale format - highest values the darkest colour and lowest values the lightest colour
- ✓ Mean values on maps are shown as a legend
- ✓ SIQ (with additional data from Agbiz Grain on the regional boundary specifications) created a software package based on an open source GIS package (QGIS)
- ✓ This software was applied again to the wheat crop quality results with great success.

ANOVA TESTS AND HOMOGENOUS GROUPS FOR MAPS

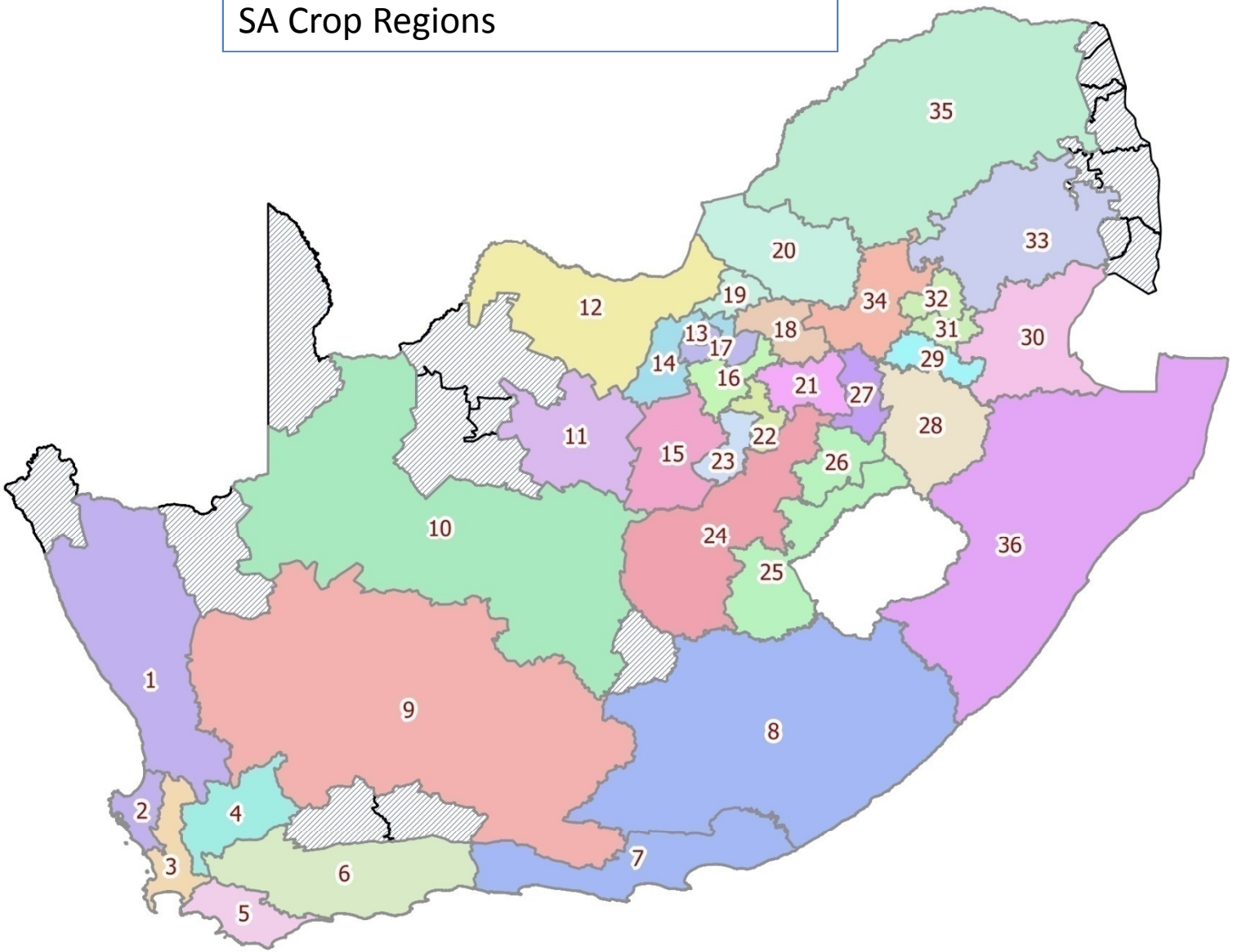
SAGL
Southern African Grain Laboratory
SINCE 1997

- ✓ Objective - identify differences between samples
- ✓ Different types of ANOVA tests – what is the question?.
- ✓ Looking for areas where specific traits may be consistently higher or lower than the average
  - ✓ 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” test
- ✓ Used Fisher LSD at 95% for the construction of the GIS maps
- ✓ GIS maps 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
- ✓ Regions without significant differences in the data were assigned the same colour on the GIS regional graph

LSD test; variable WWF falling number, sec. (kok)
Homogenous Groups, alpha = .05000 (Non-Exhaustive Search)
Error: Between MS = 3416.7, df = 2491.0

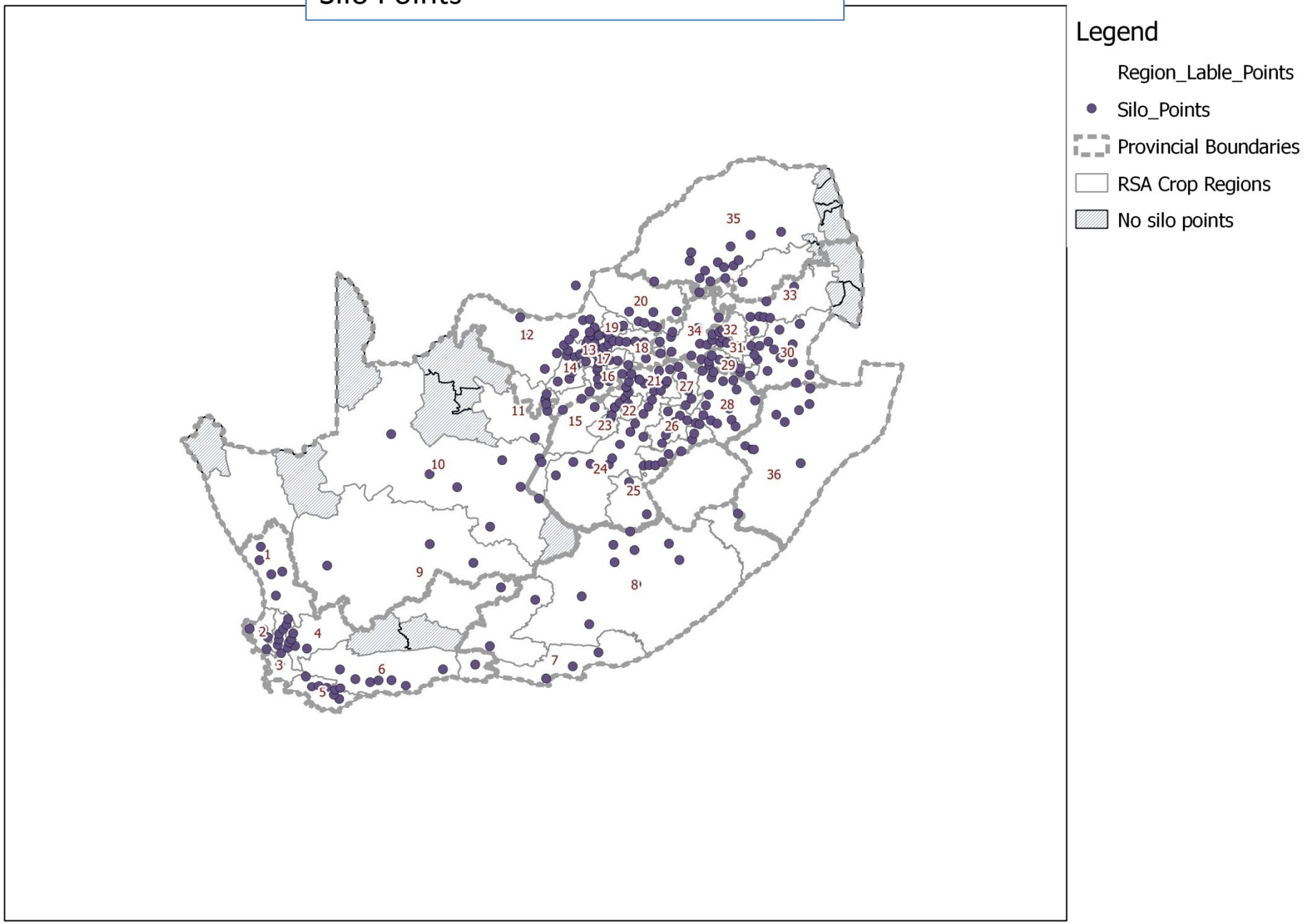
Region	WWF falling number, sec.	1	2	3	4	5	6	7	8	9	Group interval = 8.2 sec i.e. 326+8.2=334.2 etc. (400-326)/9=8.2
		326-334.2	334.2-342.4	342.4-350.6	350.6-358.8	358.8-367	367-375.2	375.2-383.4	383.4-391.6	391.6-400	
25	326	****									
32	333	****	****	***							
26	334	****	****								No samples in this group
28	338	****	****	***							
24	341		****	****							
27	342	****	****	****		****					
22	348		****	****	****	****	****				For the map the data is divided into groups based on the homogenous group interval of 8.2 seconds (95% significance)
23	349			****	****	****					
6	350			****	****	****					
21	352		****	****	****	****	****				
36	361				****		****				
17	361				****	****	****	****			
5	366						****	****			The blocks marked in yellow show three regions in one group and one region in another group. Groupings are significant but pairwise comparisons may not be significant due to high std. deviations
19	366						****	****			
15	380							****	****		
20	381								****		
4	381								****		
34	384								****		
2	384								****		
3	387								****	****	Using homogenous groups can assist with interpretations of data that will otherwise show no significant differences with pairwise ANOVA
33	387								****	****	
12	390								****	****	
11	391								****	****	
35	395								****	****	
10	400									****	

SA Crop Regions



Silo Points

QGIS



Results: KOK Wheat data

- 2003-2004 to 2014-2015 seasons
- Number: Season 3 = 2003-2004; Season 4 = 2004-2005 etc.
- Dataset very unbalanced
- Had to exclude some regions due to poor representation
- Statistics were still difficult due to large standard deviations in regional data in general, even in regions with many samples
- This data includes all seasons where data was available for the entire twelve year period
- Data stratification and sampling was done to give better balanced datasets
- Regions with too few samples can be illustrated separately if needed.

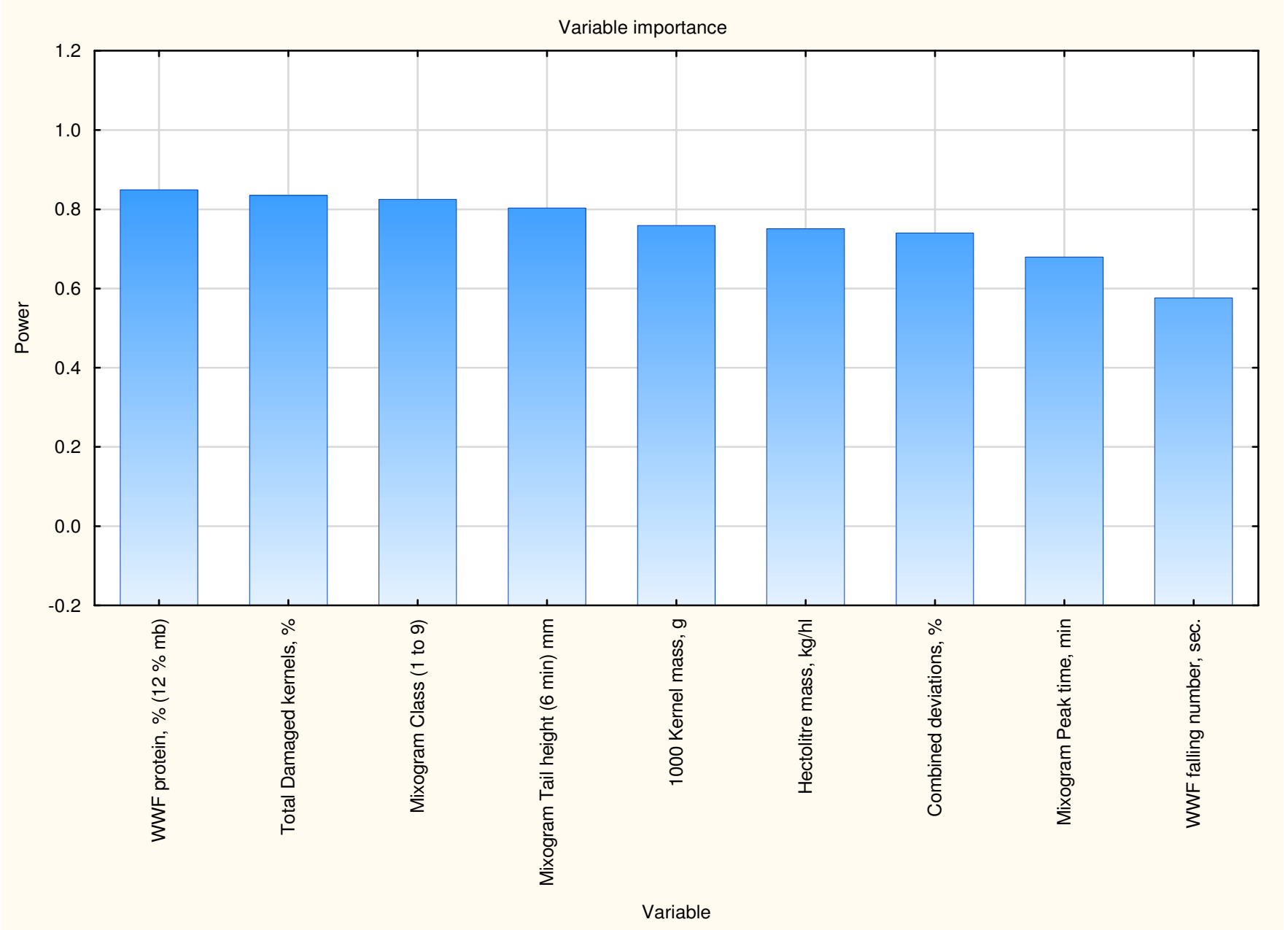
Region	Summary Stub-and-Banner Table: Frequency table for all data for KOK Marked cells have counts < 5												
	season 3	season 4	season 5	season 6	season 7	season 8	season 9	season 10	season 11	season 12	season 13	season 14	Row
1	4	3	3	0	6	4	4	3	3	0	4	4	38
2	24	19	18	18	23	24	30	12	14	20	20	14	236
3	36	62	72	65	78	71	63	44	55	69	55	51	721
4	23	51	48	17	35	14	23	25	37	28	31	31	363
5	30	40	19	27	15	19	30	20	25	19	23	17	284
6	17	21	22	33	34	34	24	11	23	35	12	19	285
7	0	1	0	0	0	0	2	1	5	0	0	0	9
8	0	0	0	0	0	4	0	0	0	0	0	0	4
10	19	16	28	27	17	23	27	32	35	31	19	23	297
11	31	11	9	14	9	24	26	14	17	16	14	12	197
12	3	0	4	4	3	7	7	5	6	2	6	4	51
14	5	5	5	3	6	0	7	4	1	1	2	4	43
15	0	6	2	13	10	9	6	9	10	3	0	0	68
16	4	0	3	1	0	0	3	3	3	0	0	0	17
17	7	3	6	4	3	6	8	8	4	1	8	2	60
18	2	4	4	0	6	3	2	2	4	0	2	2	31
19	12	12	11	11	10	13	10	8	8	2	11	2	110
20	14	28	24	25	13	25	10	15	8	2	7	15	186
21	8	10	8	12	8	2	5	5	3	1	0	1	63
22	7	6	7	3	6	10	8	6	3	4	3	3	66
23	29	15	13	17	25	23	15	22	30	14	13	15	231
24	46	16	27	27	26	17	29	16	15	7	13	21	260
25	29	24	25	39	32	31	35	25	27	18	12	19	316
26	26	26	18	18	26	25	22	13	16	6	7	6	209
27	13	8	8	8	10	3	7	8	5	6	2	3	81
28	36	29	31	33	32	29	34	31	37	21	26	15	354
29	0	1	0	0	3	0	0	1	0	1	1	1	8
30	6	4	5	4	5	3	0	1	6	6	2	0	42
32	3	3	9	7	3	7	5	1	3	0	9	7	57
33	5	17	8	11	0	10	9	0	6	2	8	6	82
34	6	5	11	17	11	18	5	11	5	8	8	8	113
35	19	26	17	22	10	17	14	8	12	13	18	28	204
36	8	8	15	0	15	5	10	8	7	1	4	4	85
Total	472	480	480	480	480	480	480	372	433	337	340	337	5171

Region	Summary Stub-and-Banner Table: Frequency table for Stratified data for KOK Marked cells have counts < 3												
	season 3	season 4	season 5	season 6	season 7	season 8	season 9	season 10	season 11	season 12	season 13	season 14	Row
2	17	11	9	7	14	14	13	6	3	10	7	9	120
3	7	6	12	13	12	8	13	9	7	13	11	9	120
4	11	14	17	6	14	2	5	8	13	12	10	8	120
5	13	14	6	14	10	10	11	9	9	8	8	8	120
6	6	9	10	12	15	10	12	5	7	17	8	9	120
10	7	7	7	8	4	11	8	14	13	17	12	12	120
11	19	6	6	8	5	14	16	7	11	10	9	9	120
12	3	0	4	4	3	7	7	5	6	2	6	4	51
15	0	6	2	13	10	9	6	9	10	3	0	0	68
17	7	3	6	4	3	6	8	8	4	1	8	2	60
19	12	12	11	11	10	13	10	8	8	2	11	2	110
20	9	17	18	18	7	17	6	10	5	1	3	9	120
21	8	10	8	12	8	2	5	5	3	1	0	1	63
22	7	6	7	3	6	10	8	6	3	4	3	3	66
23	18	8	9	12	13	15	7	7	13	7	5	6	120
24	17	6	15	10	11	8	14	12	7	5	4	11	120
25	15	8	8	16	10	14	13	9	9	6	5	7	120
26	19	19	10	11	13	12	9	9	7	4	4	3	120
27	13	8	8	8	10	3	7	8	5	6	2	3	81
28	10	8	12	16	12	9	10	10	11	7	9	6	120
32	3	3	9	7	3	7	5	1	3	0	9	7	57
33	5	17	8	11	0	10	9	0	6	2	8	6	82
34	6	5	11	17	11	18	5	11	5	8	8	8	113
35	11	13	8	16	10	8	8	6	6	7	7	20	120
36	8	8	15	0	15	5	10	8	7	1	4	4	85
Total	251	224	236	257	229	242	225	190	181	154	161	166	2516

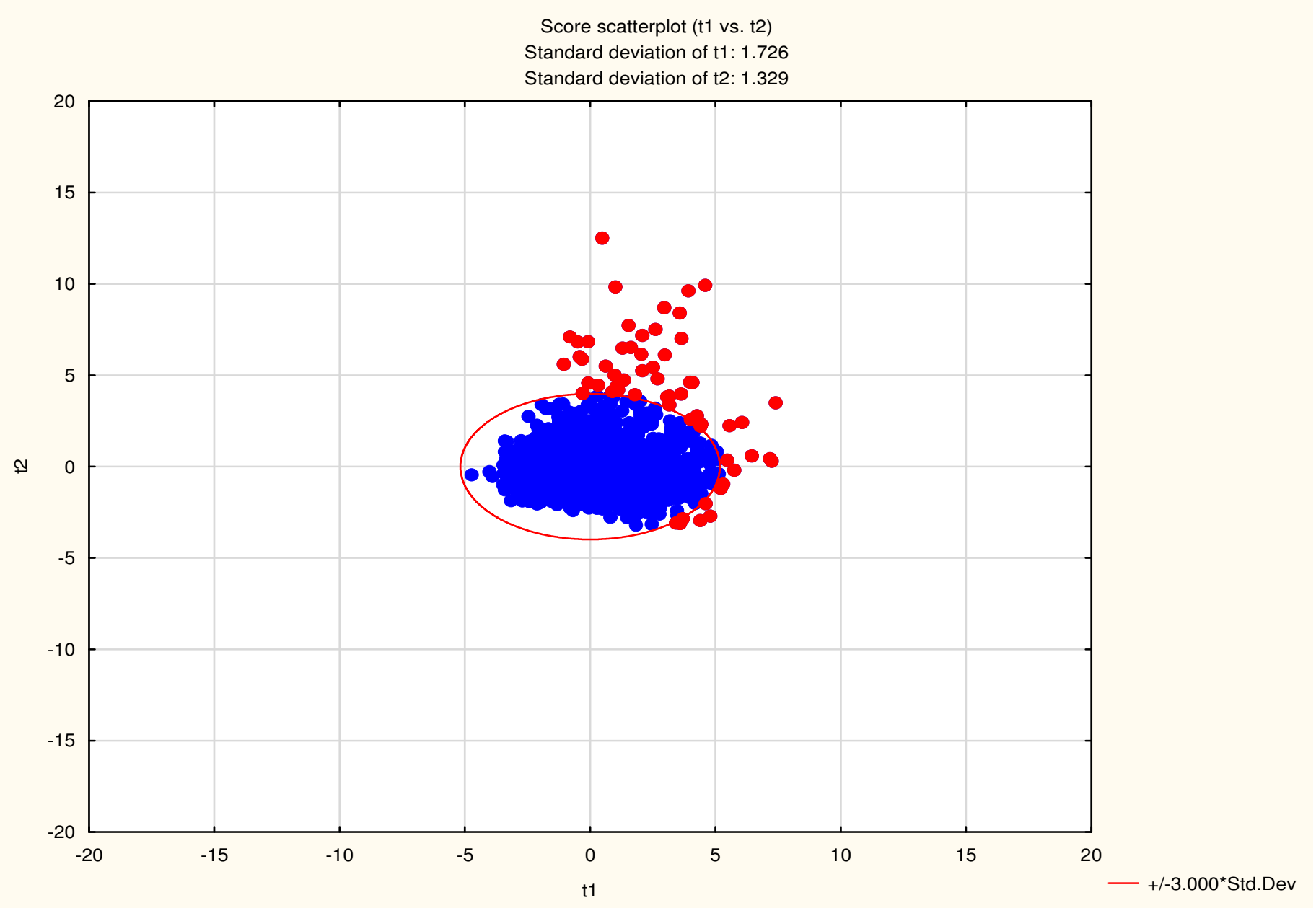
Data summaries: PCA analysis

- Principal Component Analysis (PCA) is a powerful tool based on a multivariate regression model design.
- It can summarise data in very large sets where many x-variables or predictors have been measured
- It can be visualised as a regression line fitted in a multidimensional space where each quality measurement is treated mathematically as a “dimension”
- It is used to show if a dataset is mature, whether it has subgroups, and also to pinpoint outliers
- It is also used to identify the predictor variables that are the most important i.e. those that are describing most of the data variation

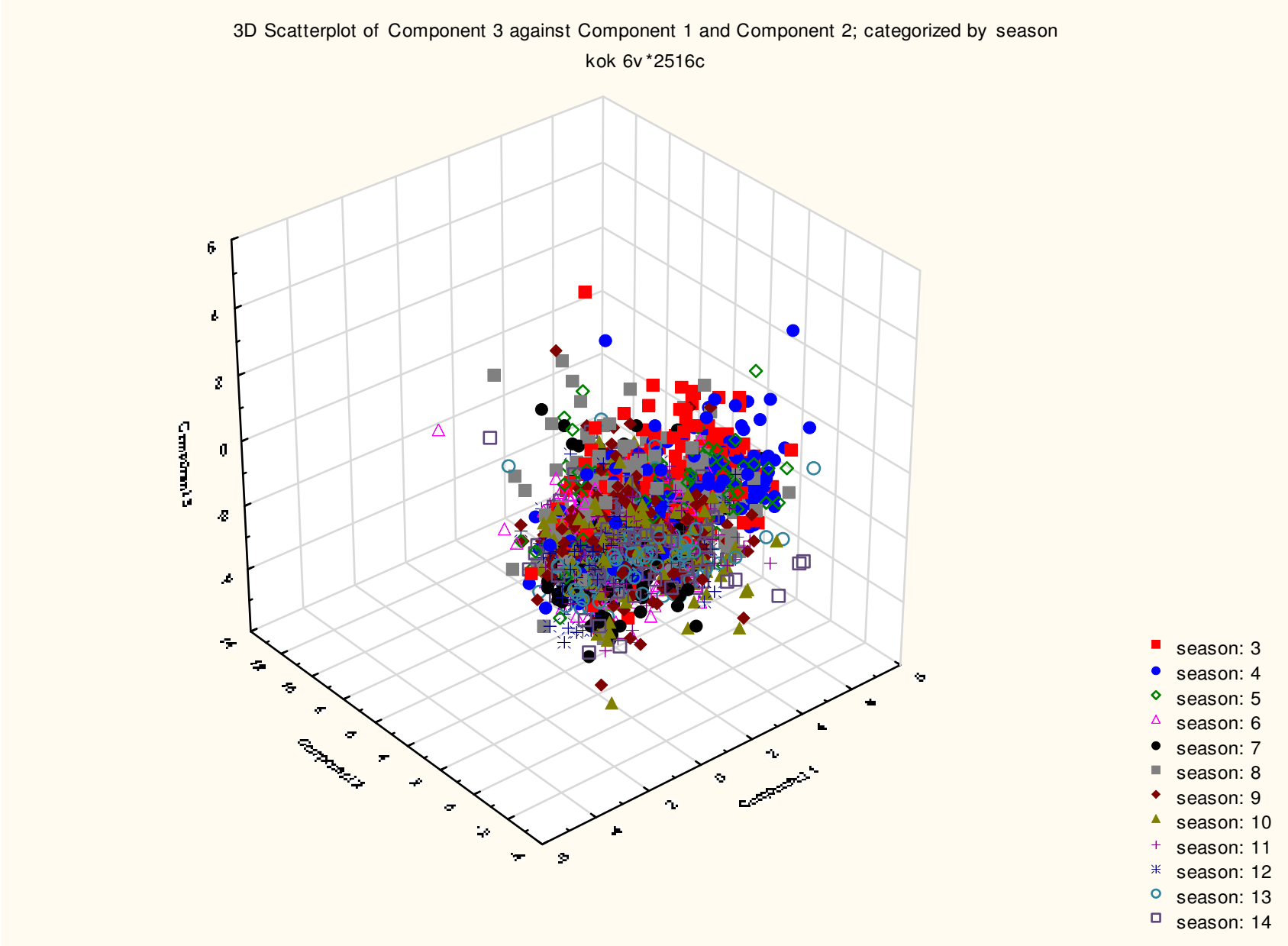
PCA Variable importance plot for KOK: summary of all data



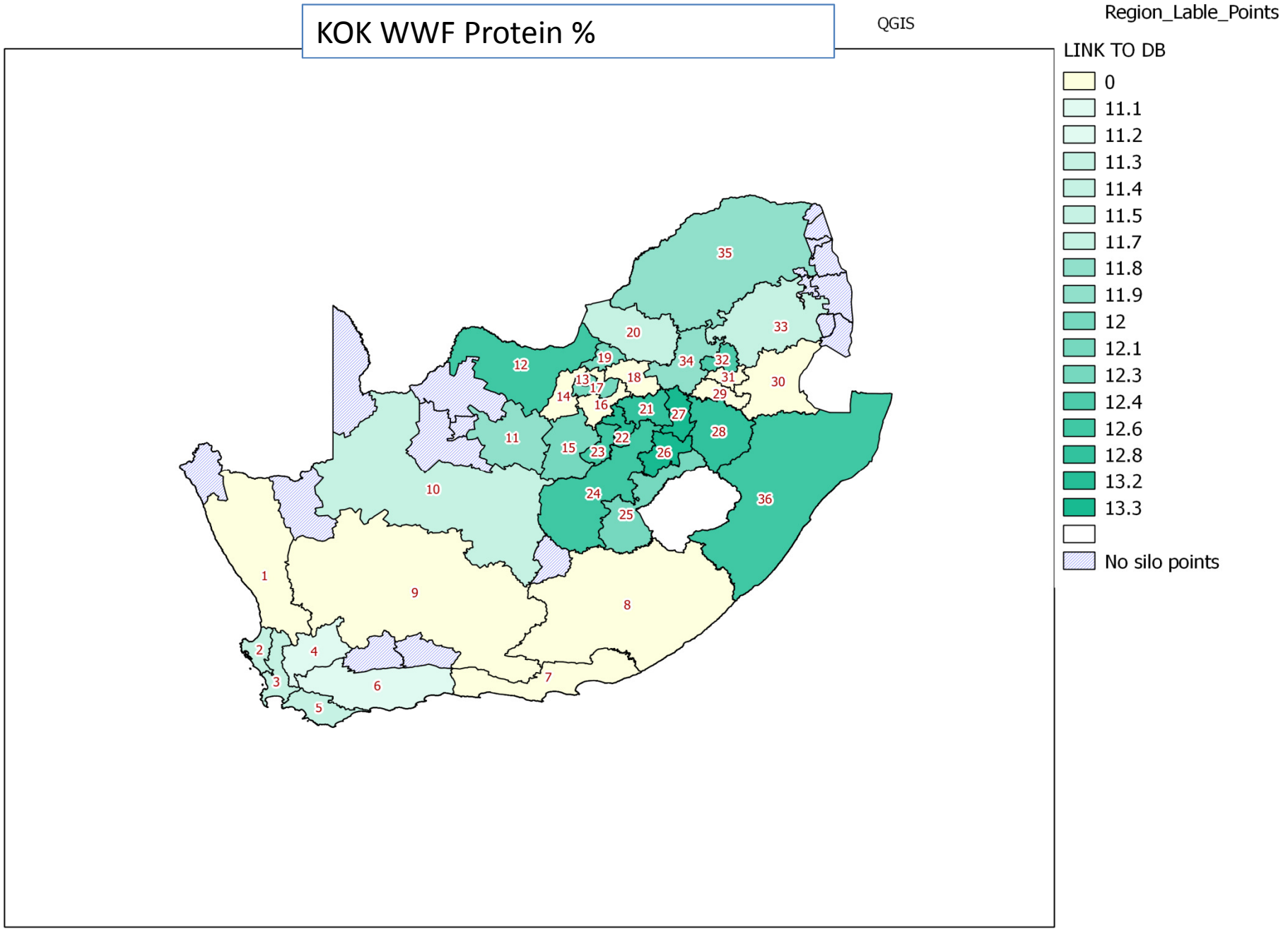
PCA Outlier analysis for KOK



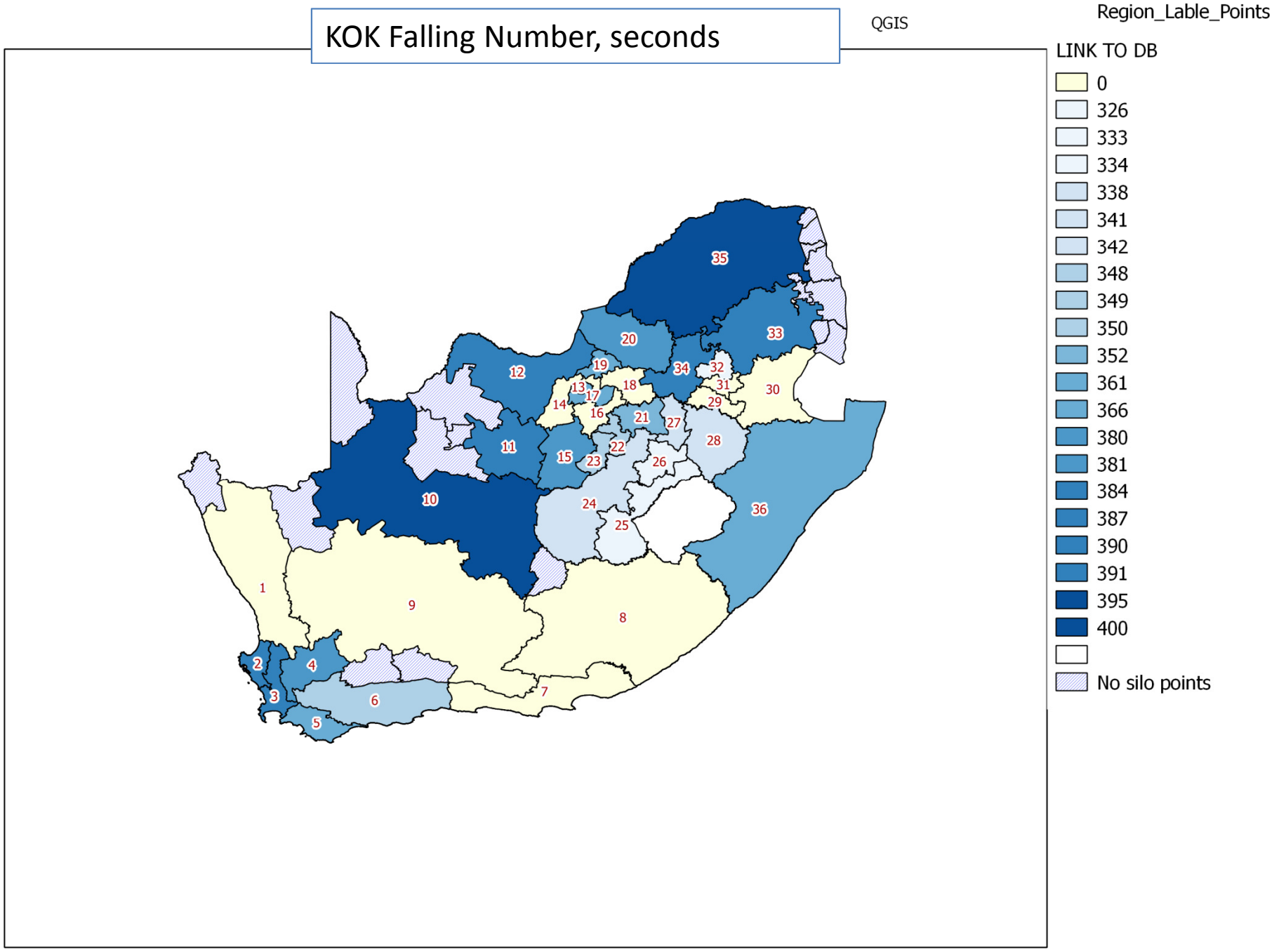
3D PCA summary plot for KOK data, classified for season



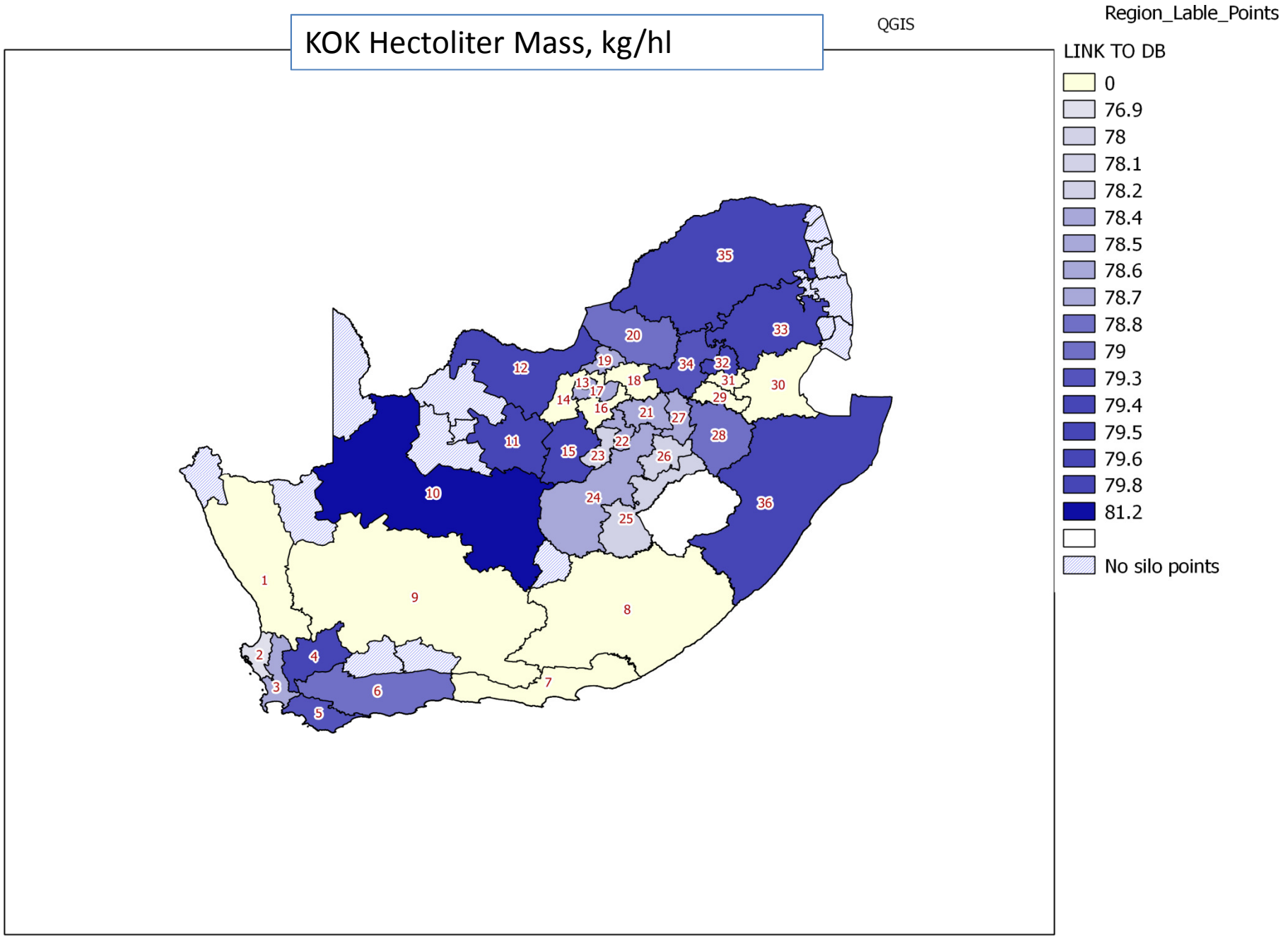
KOK WWF Protein %

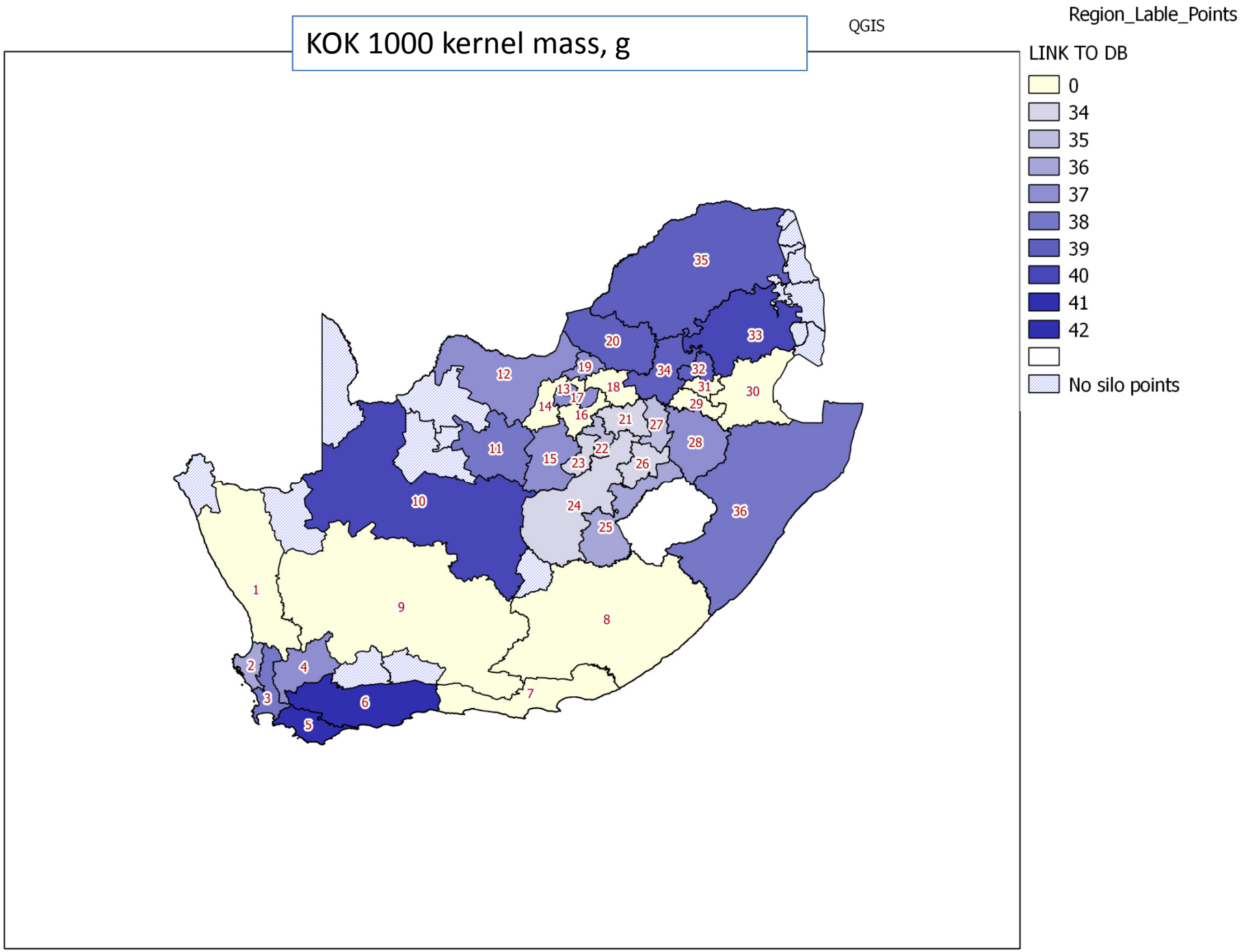


KOK Falling Number, seconds

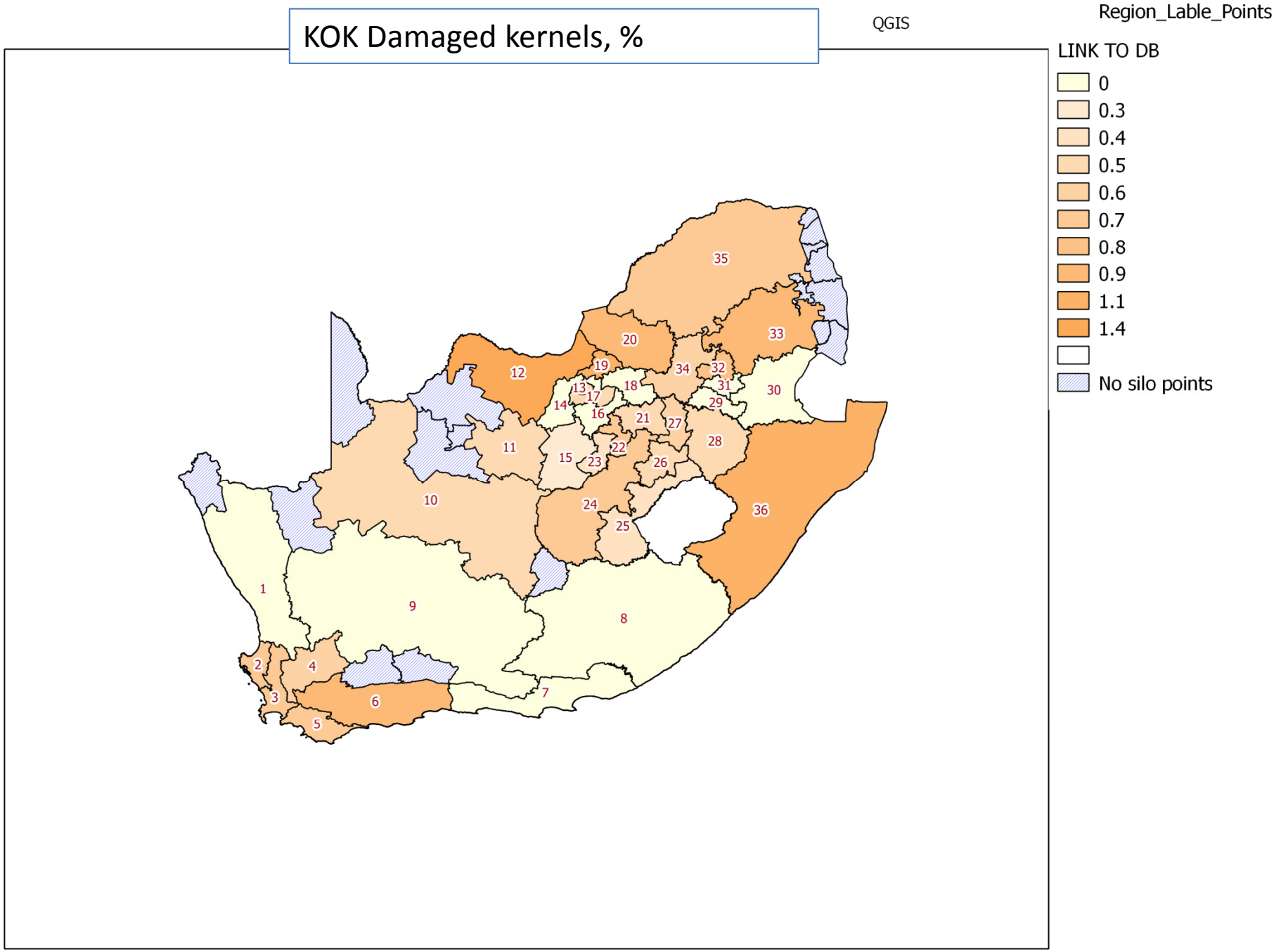


KOK Hectoliter Mass, kg/hl

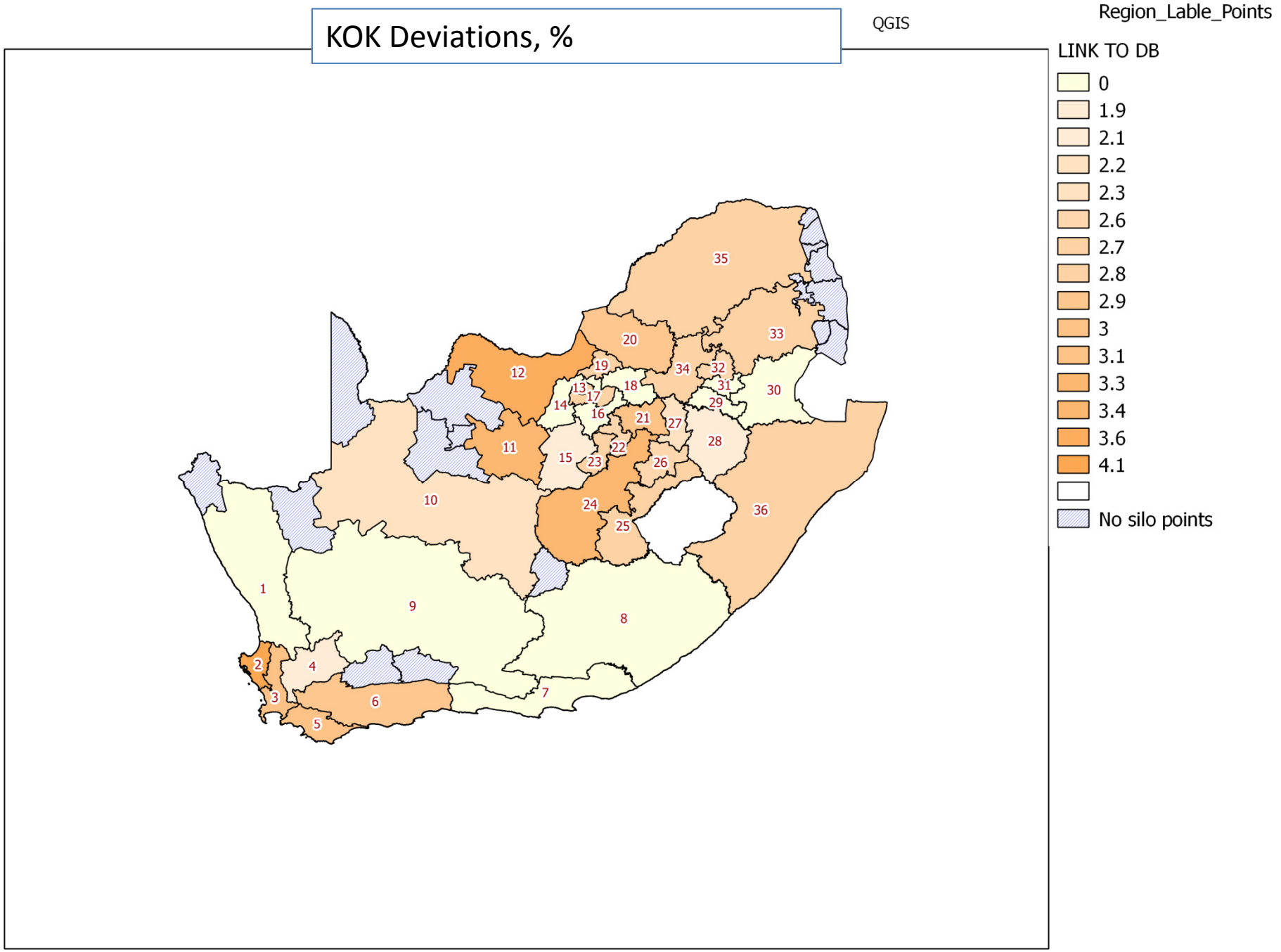



KOK 1000 kernel mass, g
QGIS
Region_Lable_Points
LINK TO DB
0
34
35
36
37
38
39
40
41
42
No silo points
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36

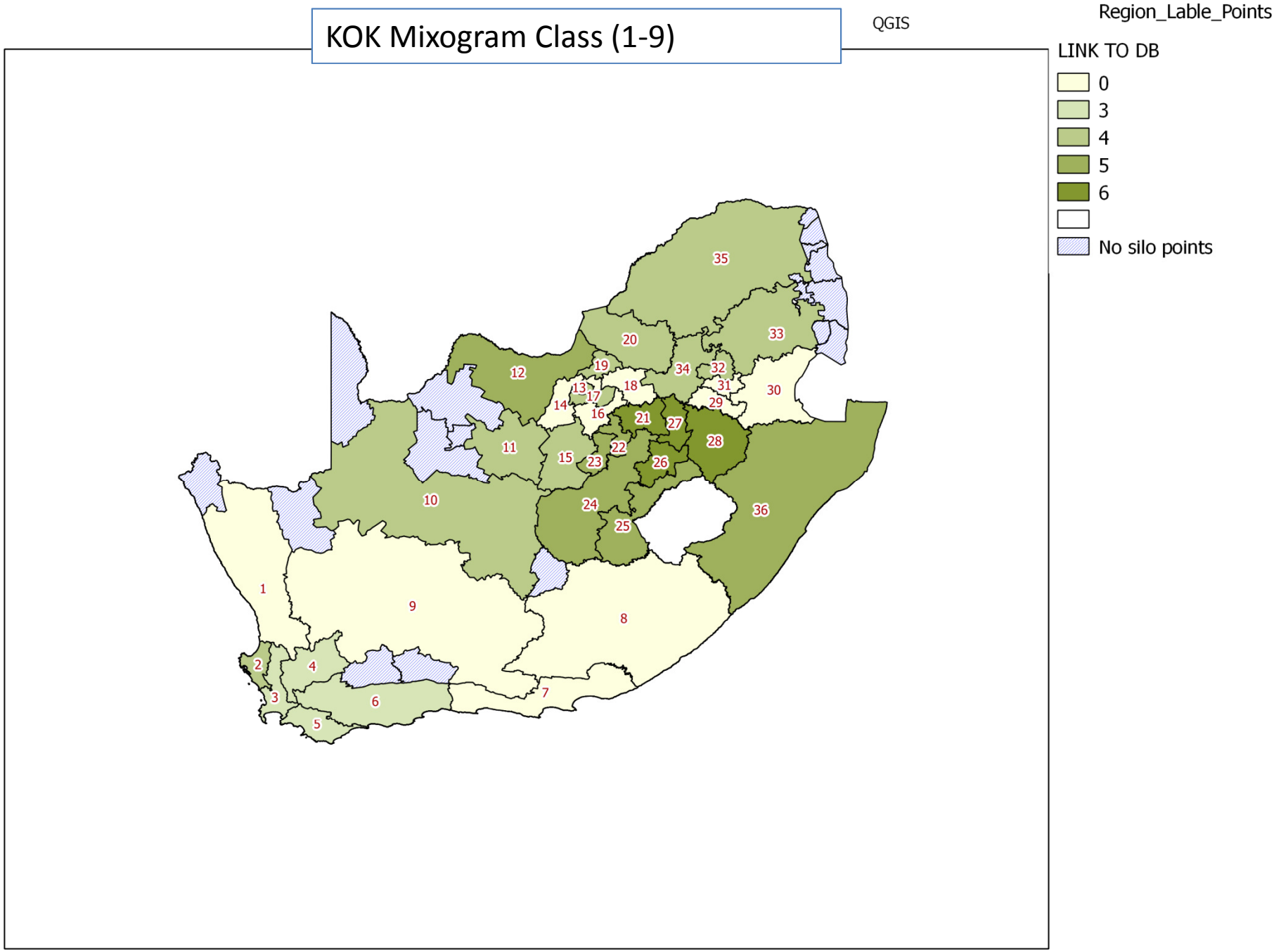
KOK Damaged kernels, %



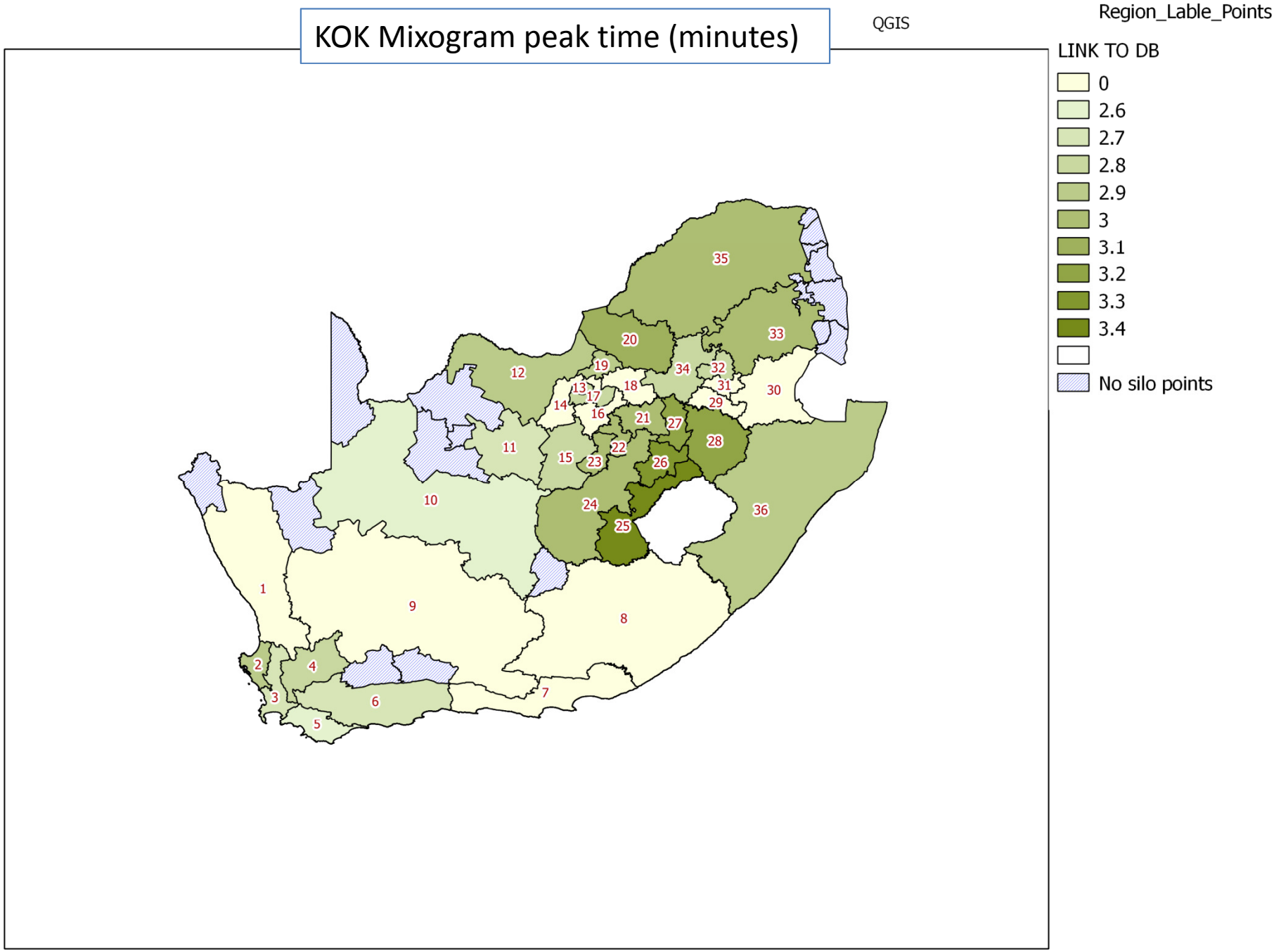
KOK Deviations, %



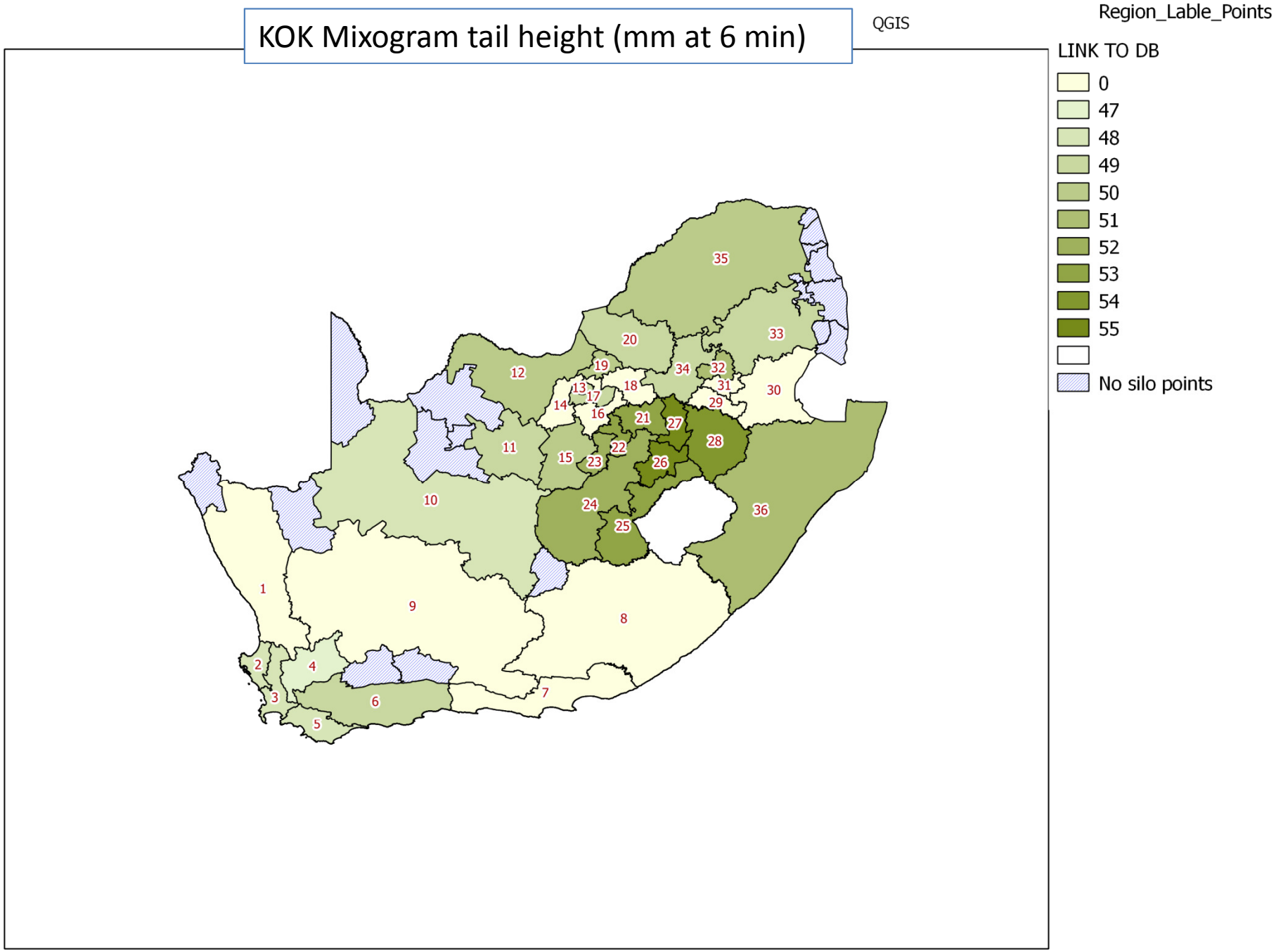
KOK Mixogram Class (1-9)



KOK Mixogram peak time (minutes)



KOK Mixogram tail height (mm at 6 min)



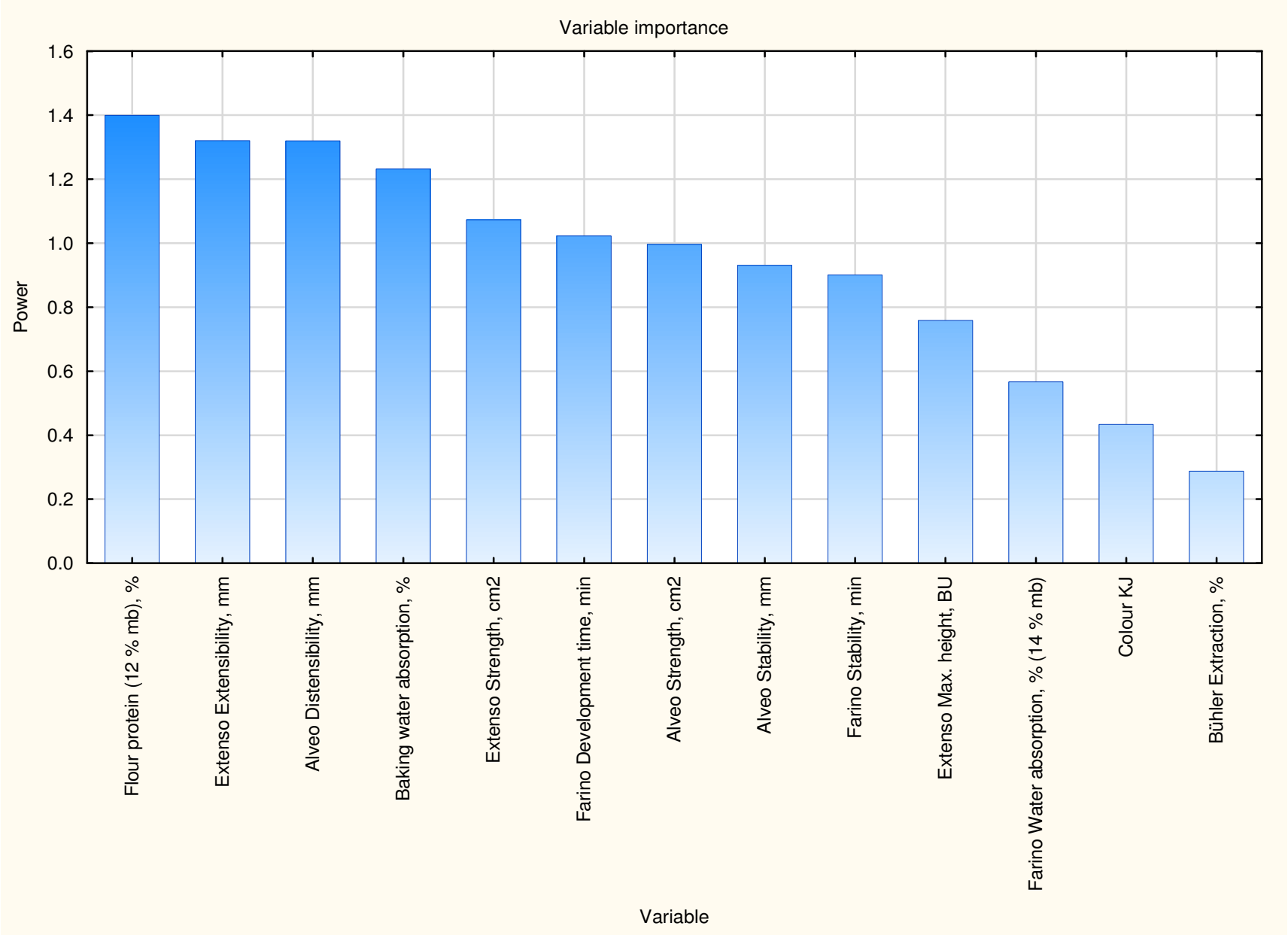
SKOK Wheat data

- 2003-2004 to 2014-2015 seasons
- Number: Season 3 = 2003-2004; Season 4 = 2004-2005 etc.
- Dataset was much smaller than KOK – the SKOK samples is a selection from the KOK samples that were analyzed more comprehensively.
- Datasets are also very unbalanced, sample stratification were done to improve the datasets.
- Had to exclude some regions due to poor representation (less than 10 samples over total period)
- The overall dataset is much smaller and with larger standard deviations than KOK samples.
- Selecting ten samples for a region as a minimum is not statistically correct for population data therefore the results must be interpreted with great care; due to many regions with relatively few samples the decision was made to work with smaller amounts of samples.

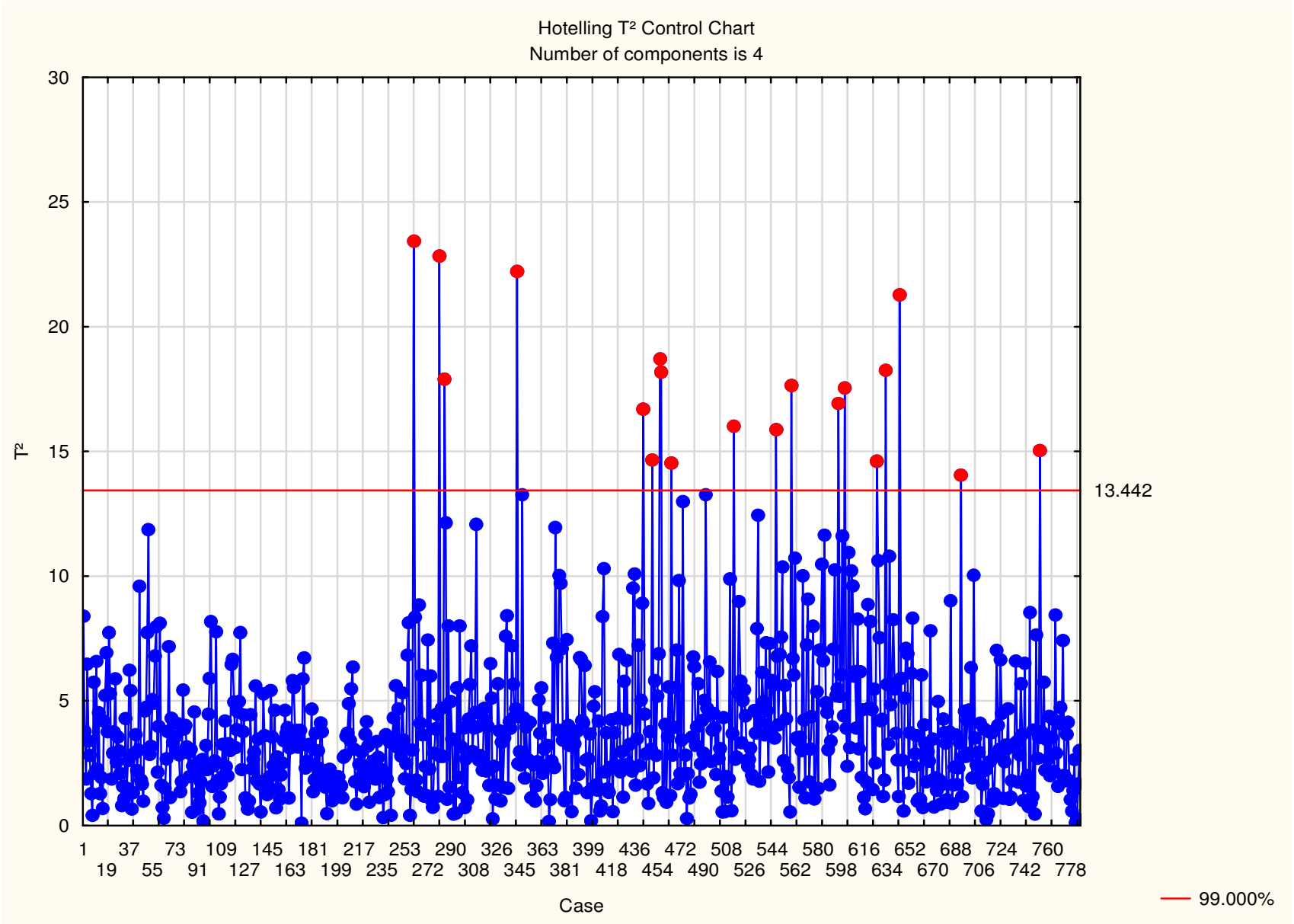
Region	Summary Stub-and-Banner Table (SKOK data for maps) Marked cells have counts <2 Season 3	Season 4	Season 5	Season 6	Season 7	Season 8	Season 9	Season 10	Season 11	Season 12	Season 13	Season 14	Row total
1	4	0	2	0	3	4	2	2	1	0	2	2	22
2	4	2	4	4	5	4	6	2	3	5	5	2	46
3	4	4	6	5	5	6	6	5	5	5	5	5	61
4	4	4	5	4	5	3	6	5	5	3	4	5	53
5	4	3	5	5	4	5	5	5	3	4	4	3	50
6	4	4	4	5	5	5	4	2	5	3	3	4	48
7	0	1	0	0	0	0	0	1	1	0	0	0	3
8	0	0	0	0	0	1	0	0	0	0	0	0	1
10	4	3	5	3	4	3	3	5	4	4	4	4	46
11	4	3	4	4	2	4	4	4	4	4	4	3	44
12	0	0	1	2	1	2	3	2	1	1	1	1	15
14	1	1	3	2	3	0	3	2	1	0	1	1	18
15	0	3	1	4	4	3	3	3	4	3	0	0	28
16	0	0	1	1	0	0	2	2	1	0	0	0	7
17	4	2	2	3	1	2	3	3	4	1	2	1	28
18	1	1	2	0	2	1	0	2	1	0	1	1	12
19	3	3	4	3	3	4	3	3	2	1	1	1	31
20	3	4	5	6	4	4	3	5	3	1	3	2	43
21	2	2	2	4	2	2	3	3	2	1	0	1	24
22	2	2	3	2	2	3	3	4	2	2	1	1	27
23	4	4	3	4	6	4	4	4	6	2	3	4	48
24	4	3	4	5	5	5	5	5	4	2	3	4	49
25	4	3	4	5	6	4	5	6	4	5	3	5	54
26	4	4	5	4	5	5	4	3	3	2	2	2	43
27	3	3	3	2	3	3	3	3	2	3	1	1	30
28	4	4	4	3	4	3	3	6	3	4	4	3	45
29	0	1	0	0	1	0	0	1	0	1	0	0	4
30	1	3	2	3	1	1	0	1	1	2	0	0	15
32	0	1	2	3	2	2	3	1	2	0	2	2	20
33	2	3	3	5	0	4	3	0	4	1	2	2	29
34	2	2	2	3	4	5	2	4	4	4	3	3	38
35	3	4	5	6	4	5	3	3	3	5	4	5	50
36	1	3	4	0	4	3	2	1	1	1	2	2	24
Total	80	80	100	100	100	100	99	98	89	70	70	70	1056

Region	Summary Stub-and-Banner Table (SKOK data for maps), stratified balanced sheet Marked cells have counts <2 Season 3	Season 4	Season 5	Season 6	Season 7	Season 8	Season 9	Season 10	Season 11	Season 12	Season 13	Season 14	Row total
1	4	0	2	0	3	4	2	2	1	0	2	2	22
2	4	0	3	4	5	3	4	2	1	3	3	2	34
3	2	3	3	2	2	3	5	1	1	2	1	4	29
4	3	3	3	3	4	1	3	4	4	3	1	4	36
5	3	1	2	2	3	4	2	5	3	1	2	1	29
6	1	2	4	4	3	3	4	0	4	0	2	3	30
10	3	3	2	2	3	1	3	4	3	2	4	2	32
11	3	3	3	3	1	3	3	4	3	4	3	1	34
12	0	0	1	2	1	2	3	2	1	1	1	1	15
14	1	1	3	2	3	0	3	2	1	0	1	1	18
15	0	3	1	4	4	3	3	3	4	3	0	0	28
17	4	2	2	3	1	2	3	3	4	1	2	1	28
18	1	1	2	0	2	1	0	2	1	0	1	1	12
19	3	3	4	3	3	4	3	3	2	1	1	1	31
20	2	2	4	6	4	4	3	4	1	1	1	2	34
21	2	2	2	4	2	2	3	3	2	1	0	1	24
22	2	2	3	2	2	3	3	4	2	2	1	1	27
23	2	4	0	2	5	3	4	3	2	1	3	3	32
24	2	1	3	2	2	3	4	2	3	2	3	2	29
25	3	1	2	2	2	0	2	3	2	4	2	1	24
26	1	3	3	4	4	4	3	2	2	2	1	1	30
27	3	3	3	2	3	3	3	3	2	3	1	1	30
28	3	2	2	3	3	2	1	4	1	2	0	1	24
30	1	3	2	3	1	1	0	1	1	2	0	0	15
32	0	1	2	3	2	2	3	1	2	0	2	2	20
33	2	3	3	5	0	4	3	0	4	1	2	2	29
34	2	2	2	1	4	4	2	3	4	4	3	2	33
35	1	3	2	4	4	3	1	2	1	4	2	2	29
36	1	3	4	0	4	3	3	2	2	1	2	2	27
Total	59	60	72	77	80	75	79	74	64	51	47	47	785

SKOK PLS Regression: loaf volume



Outliers: SKOK PLS Regression, loaf volume

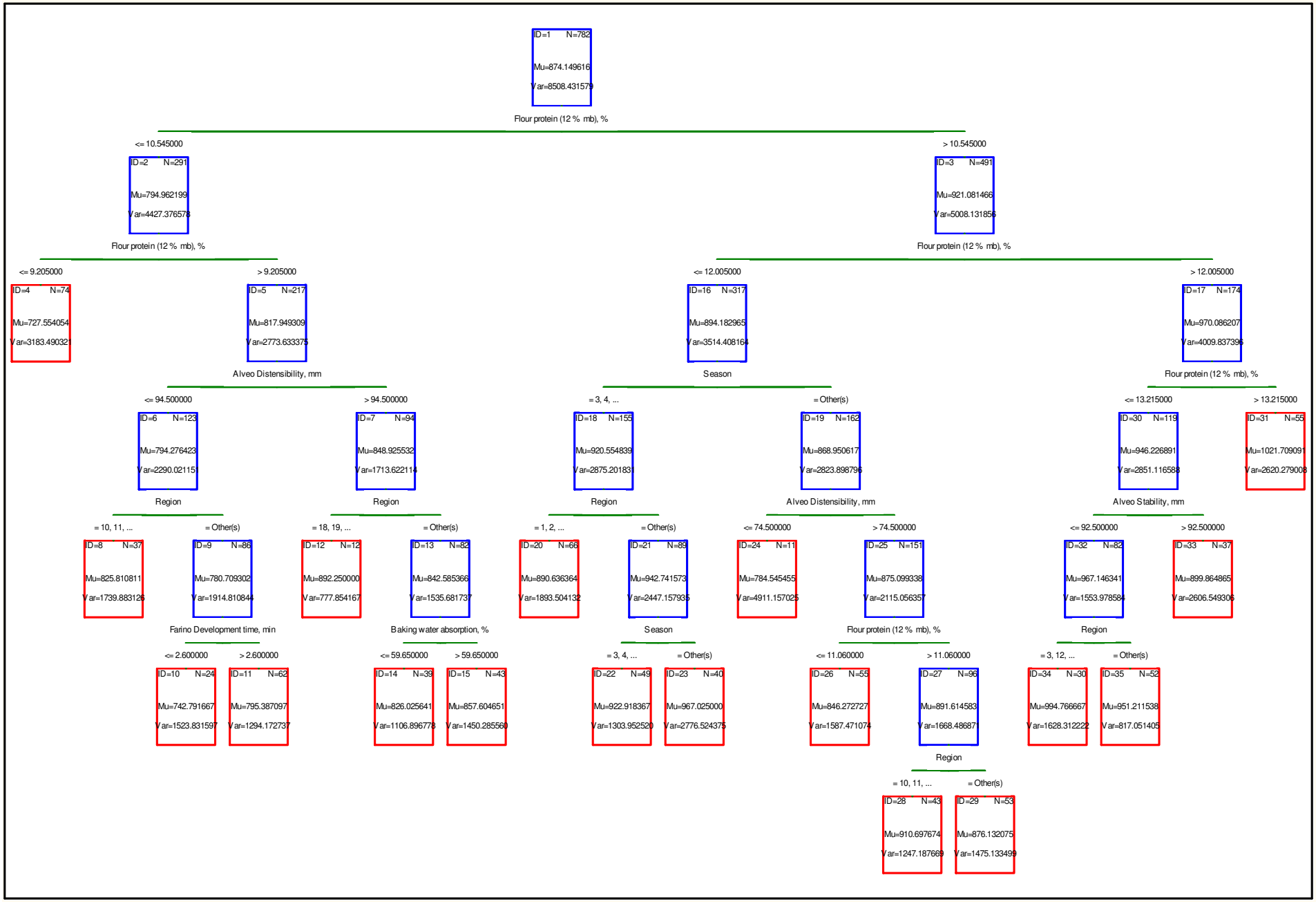


REGRESSION TREES

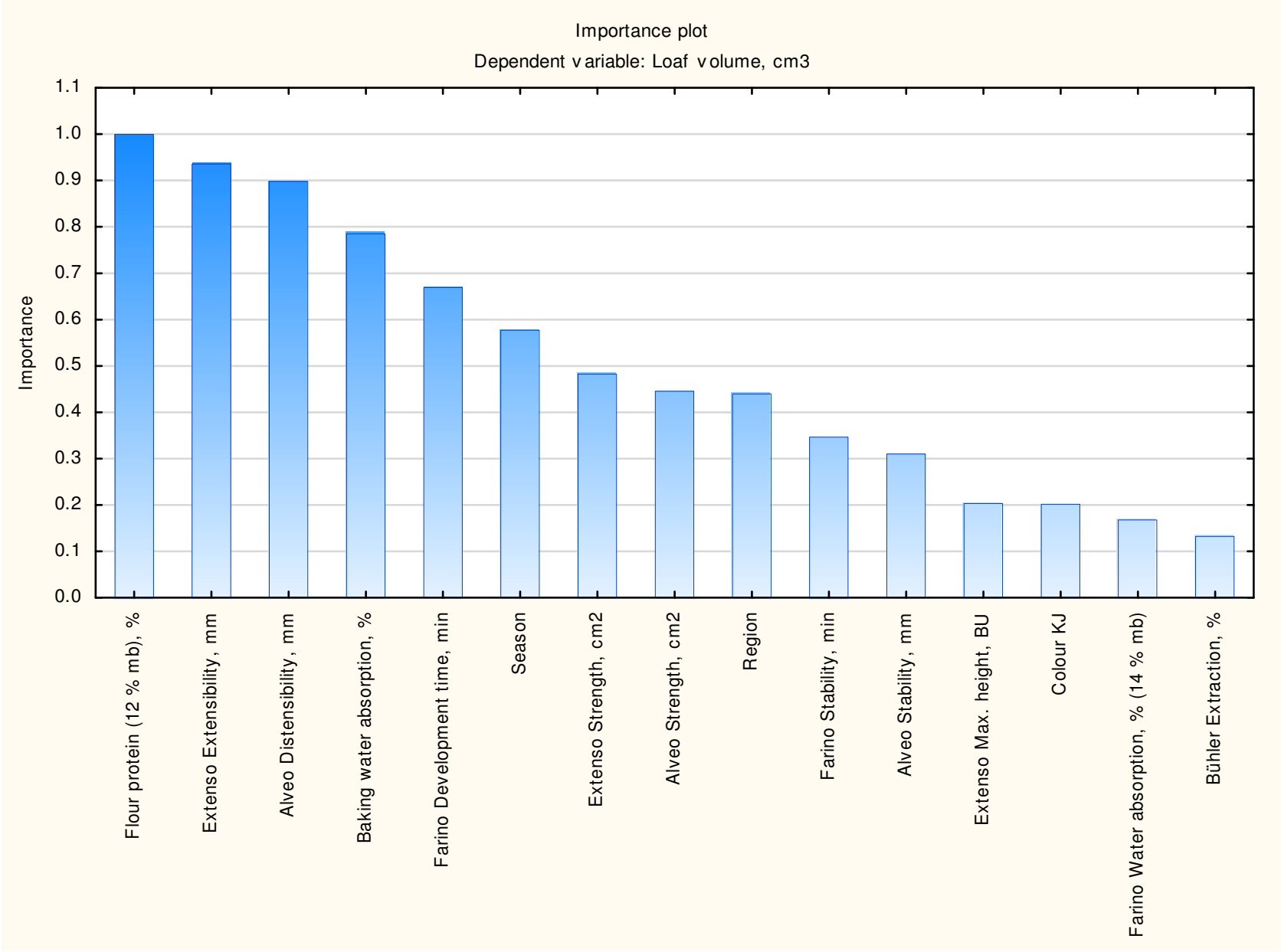
- ✓Regression tree example shows the effects of all variables (factors and continuous) on loaf volume for the twelve year period
- ✓785 samples were included in the model (the stratified balanced dataset)
- ✓Trees are very complex to show visually, but the results are summarised again in a variable importance plot for ease of use (Statistica 13 software from Dell).

Tree 1 graph for Loaf volume, cm3

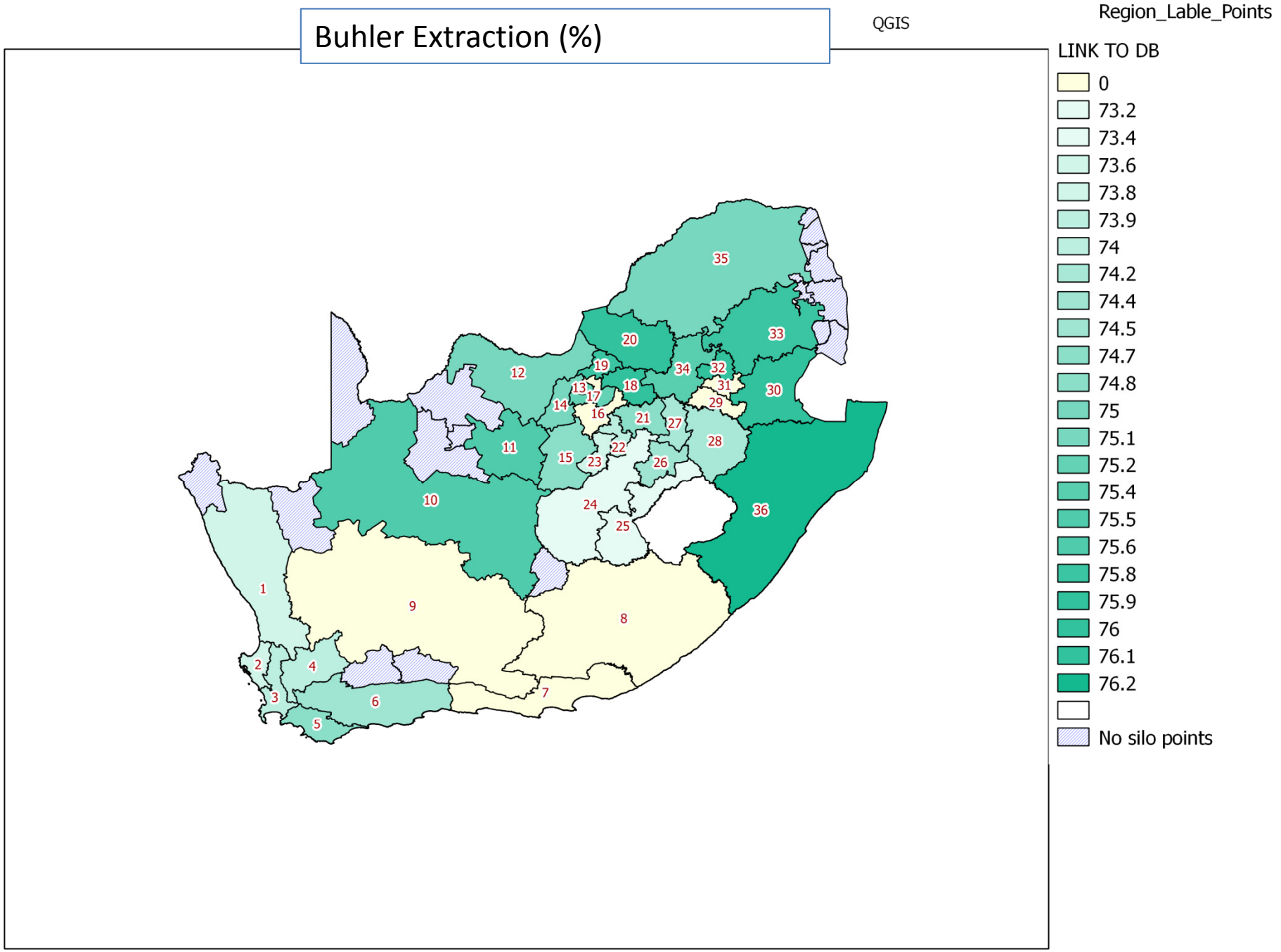
Num. of non-terminal nodes: 17, Num. of terminal nodes: 18



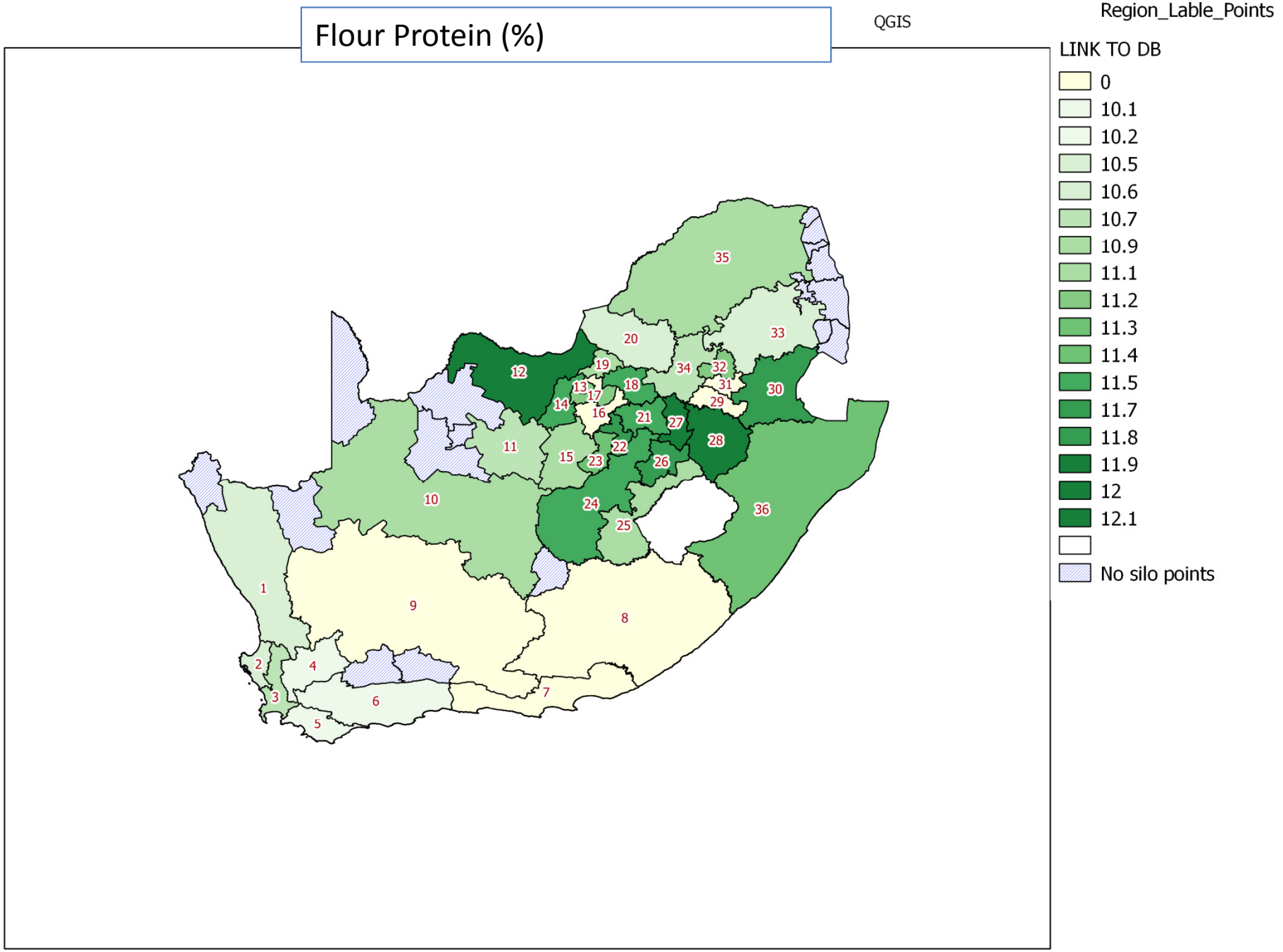
Regression tree SKOK: loaf volume



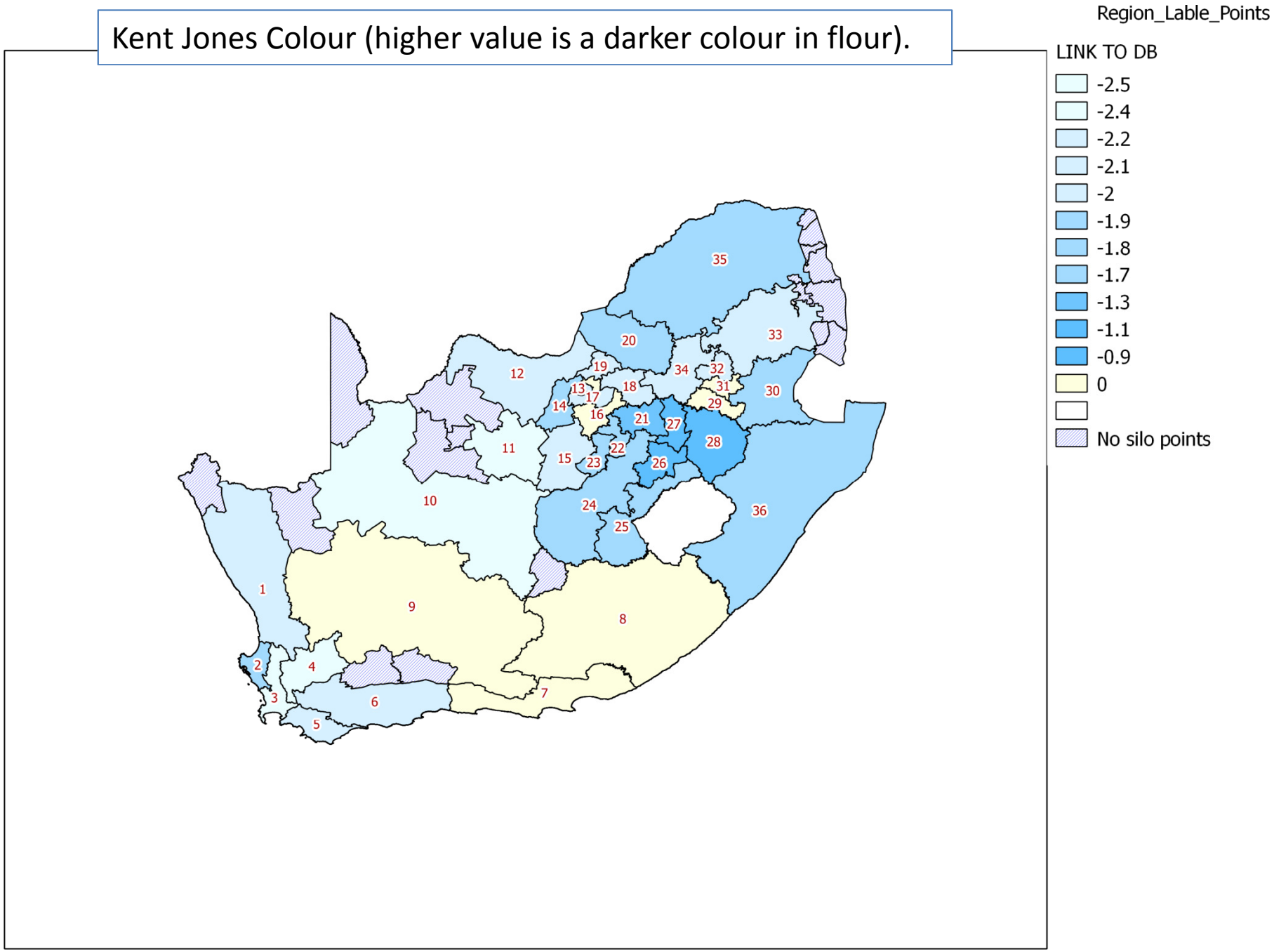
Buhler Extraction (%)



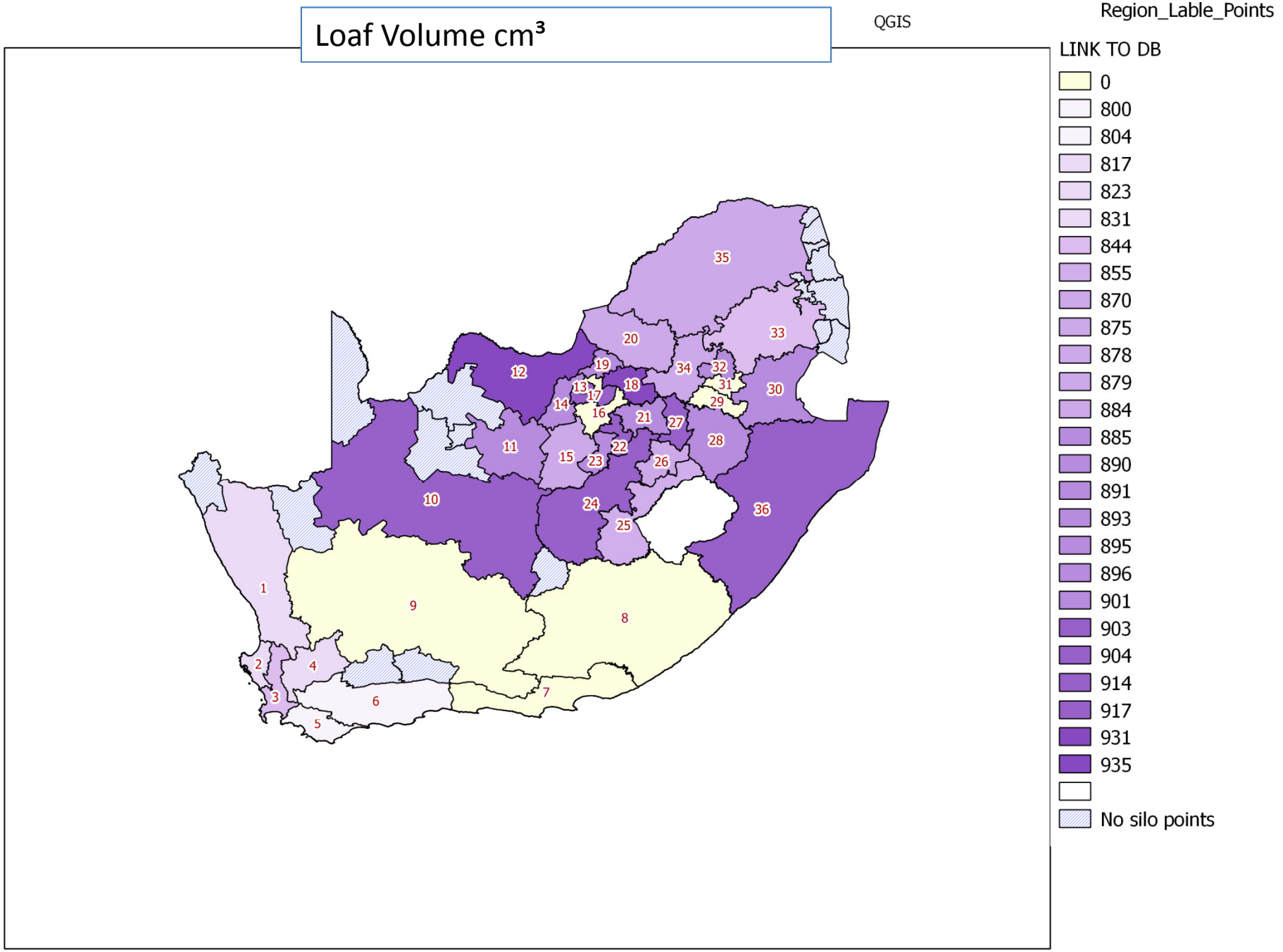
Flour Protein (%)



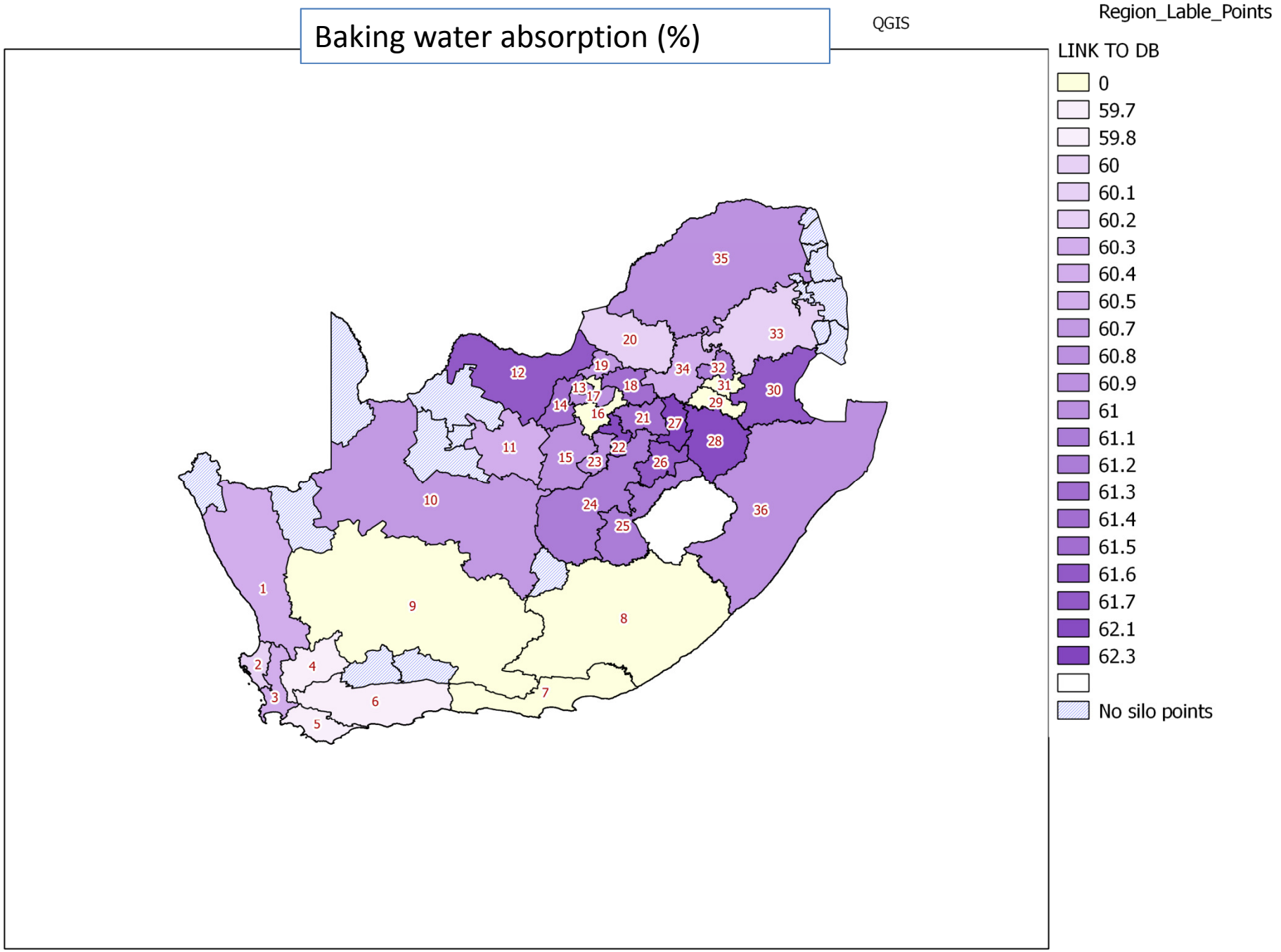
Kent Jones Colour (higher value is a darker colour in flour).



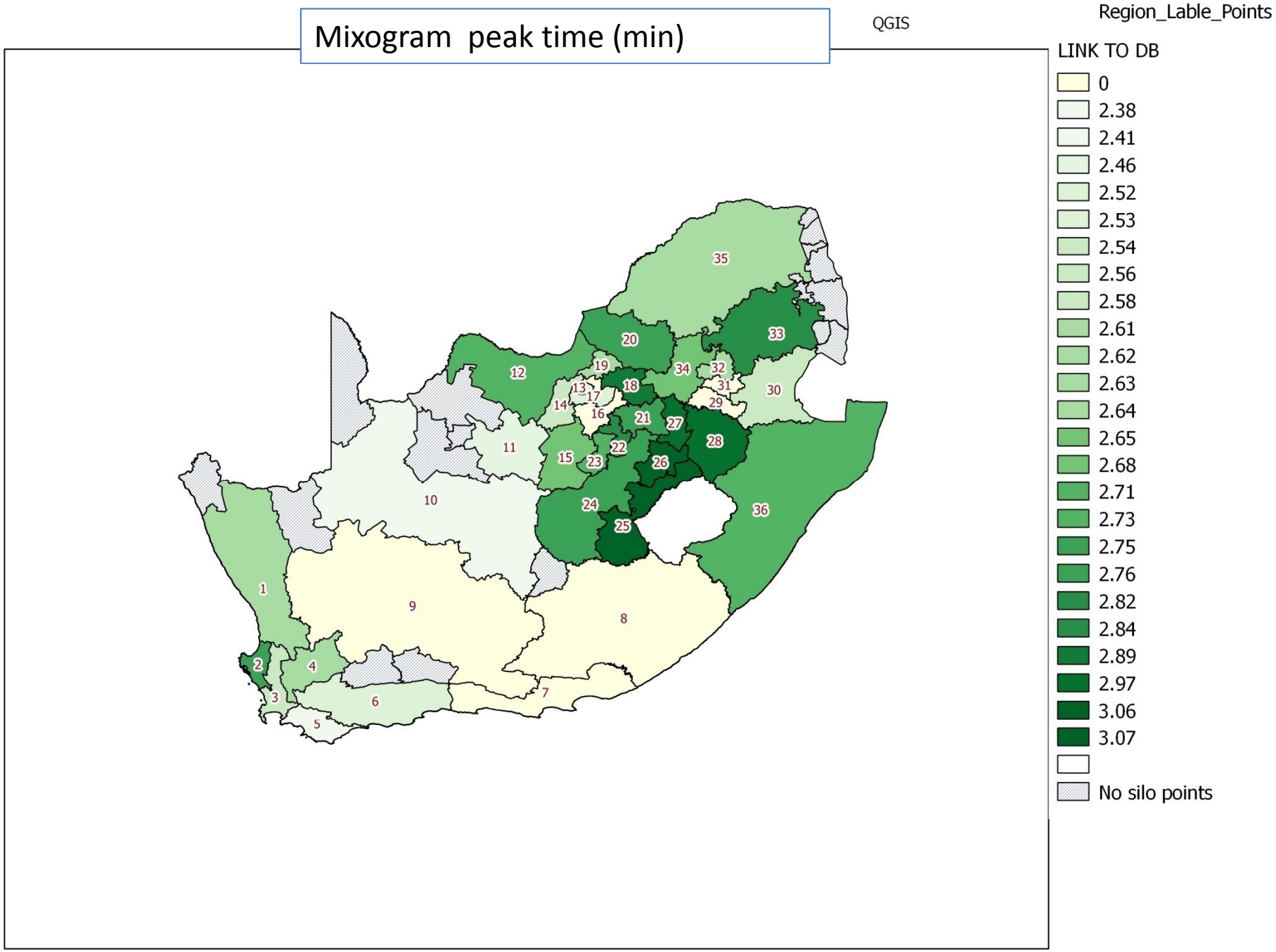
Loaf Volume cm³



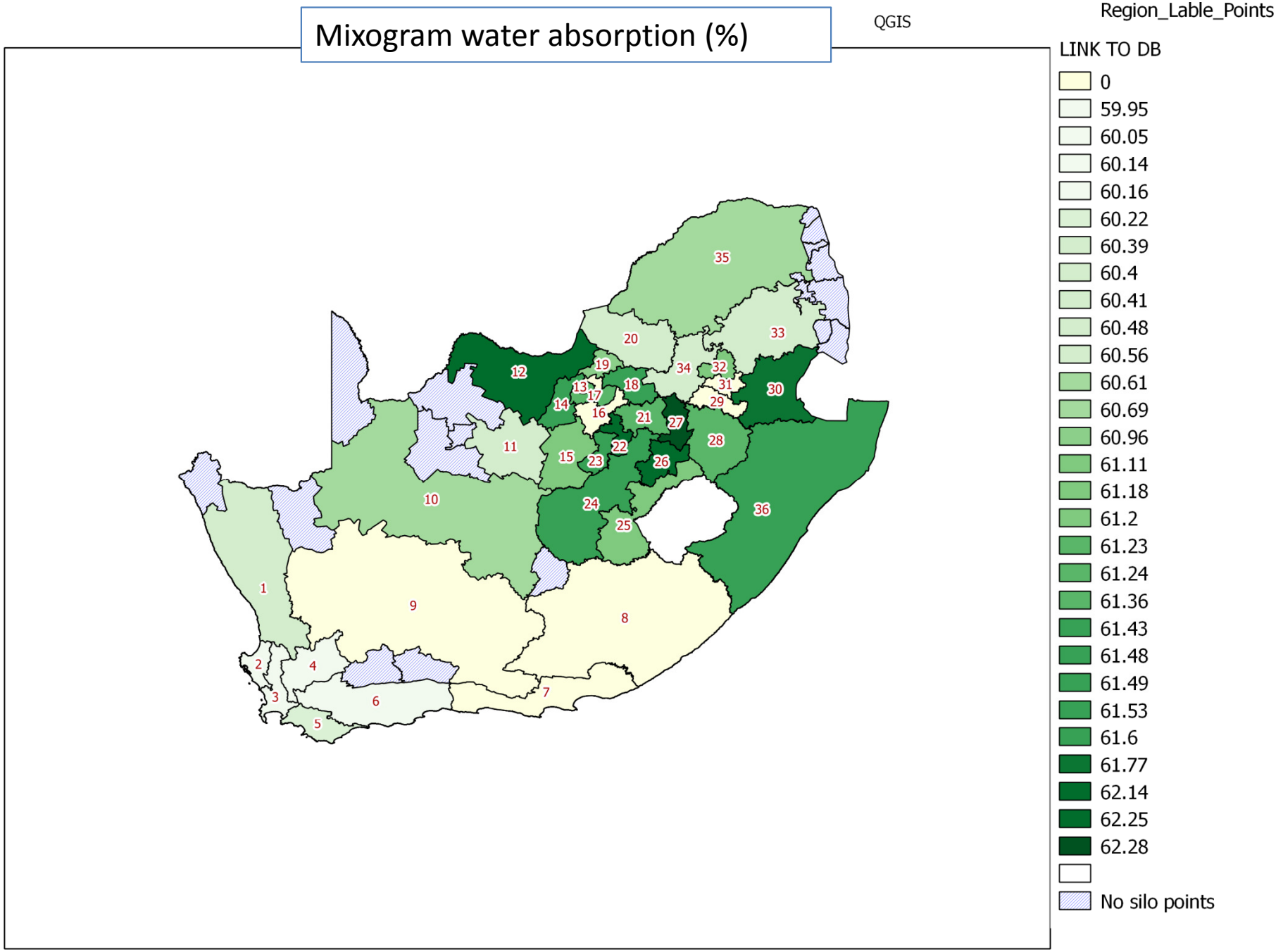
Baking water absorption (%)



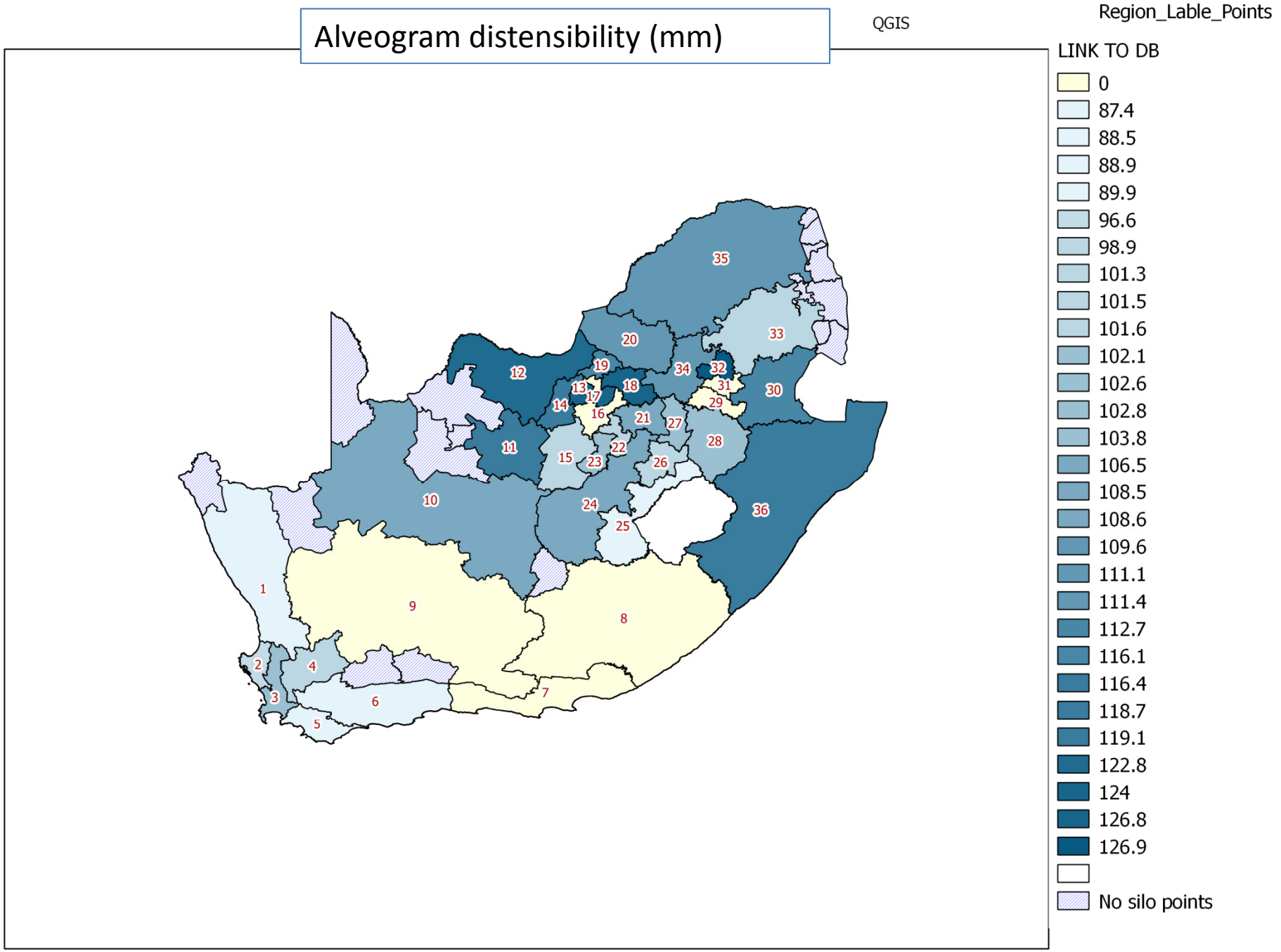
Mixogram peak time (min)



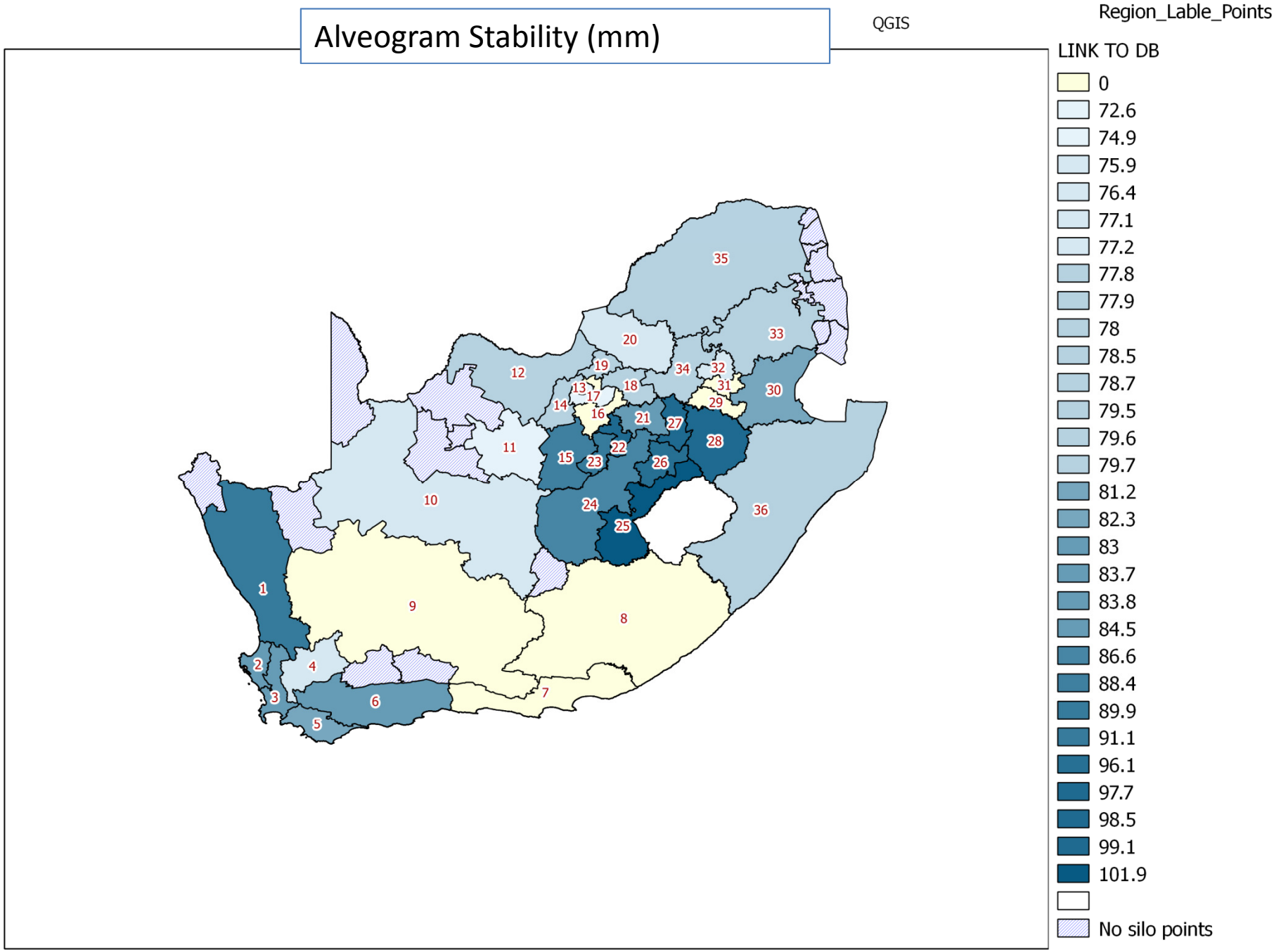
Mixogram water absorption (%)



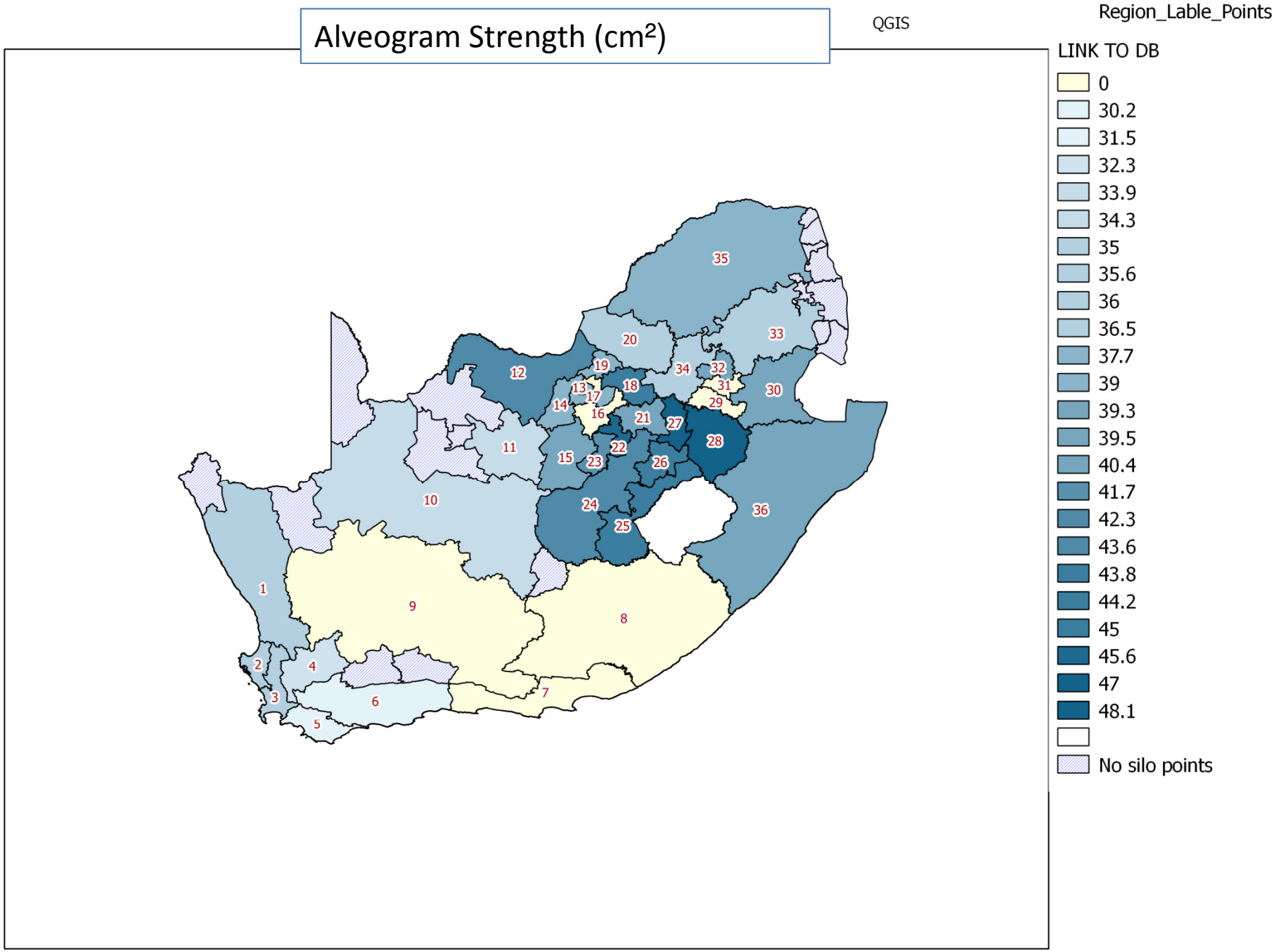
Alveogram distensibility (mm)



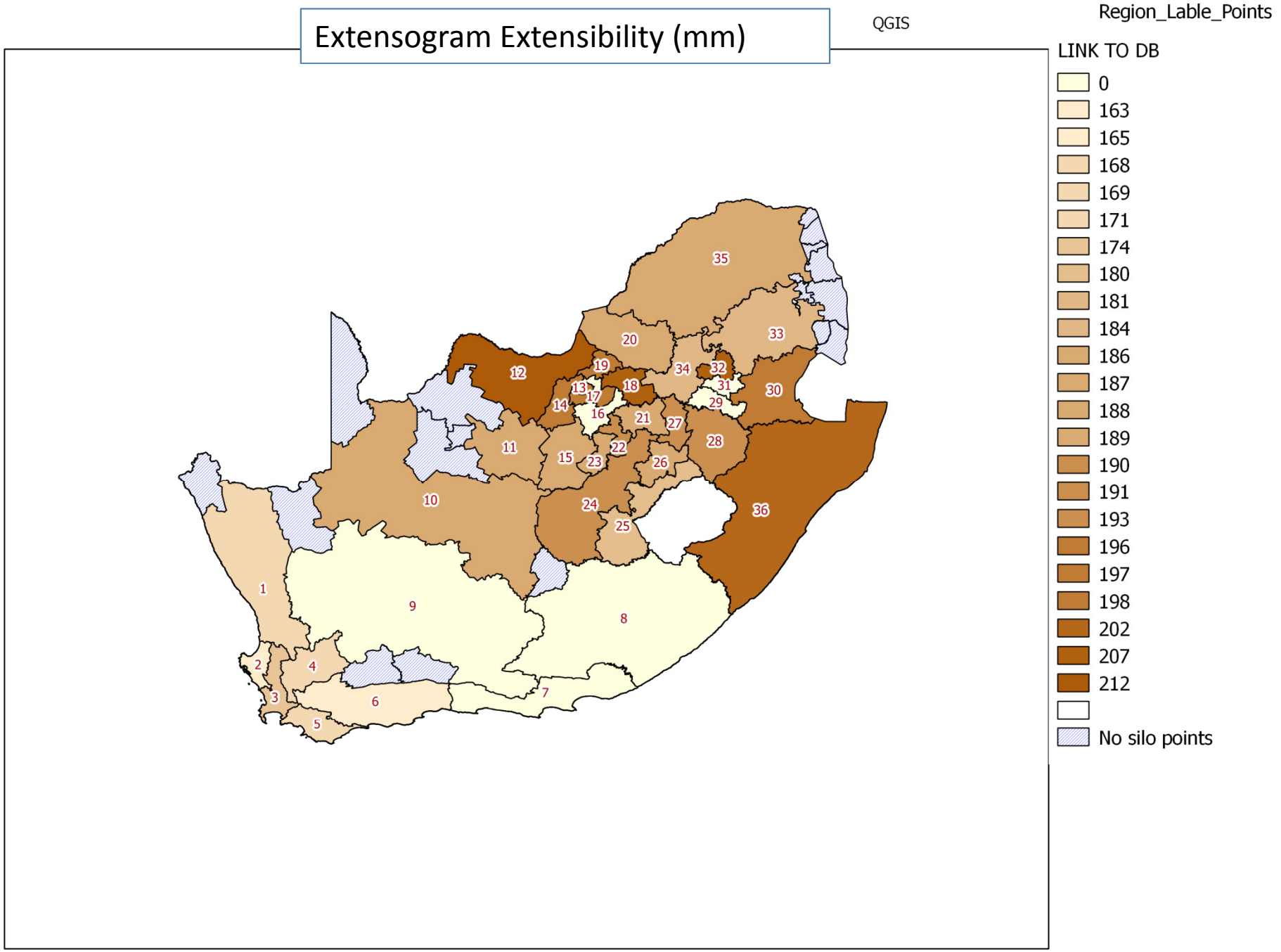
Alveogram Stability (mm)



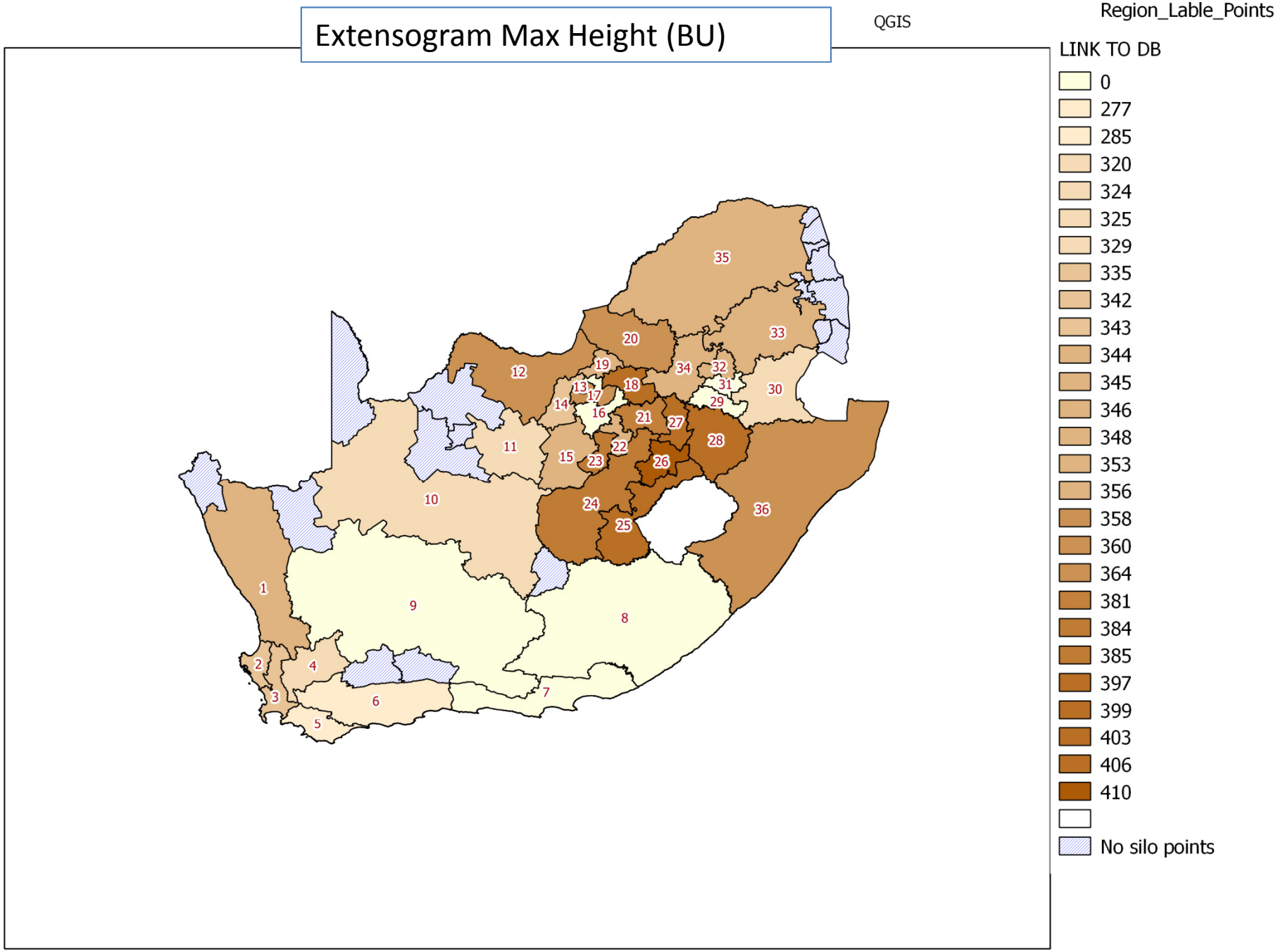
Alveogram Strength (cm²)



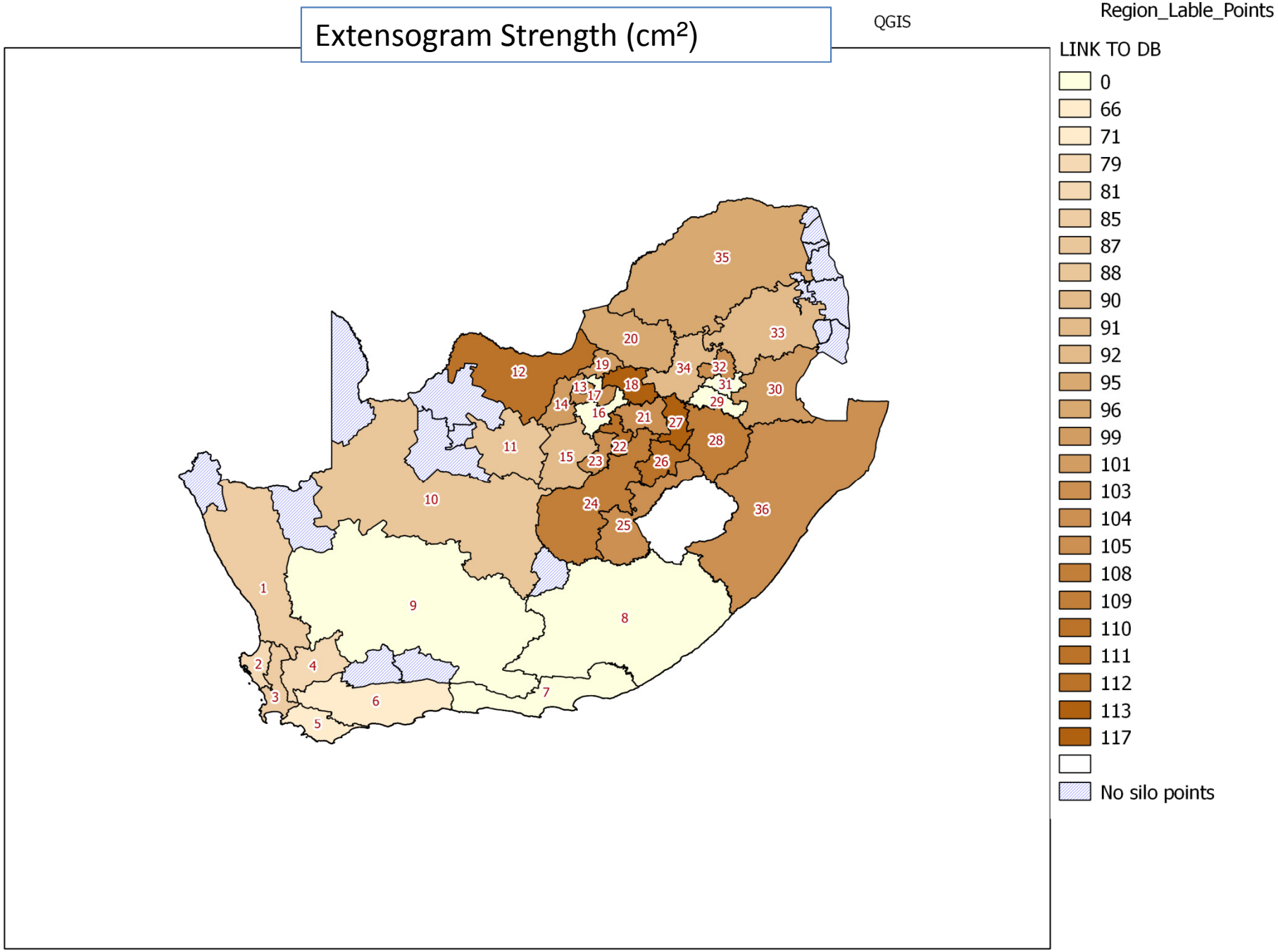
Extensogram Extensibility (mm)



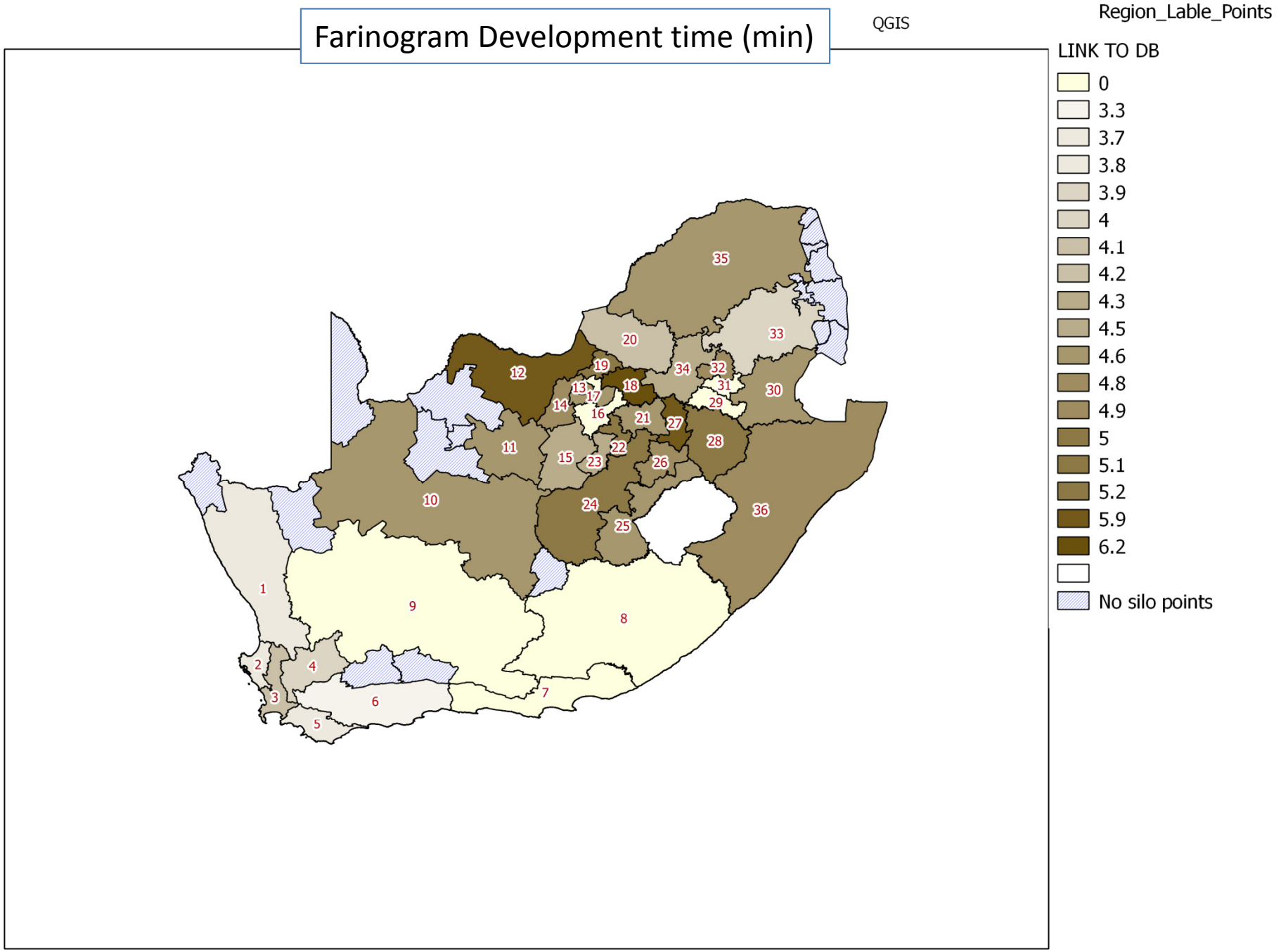
Extensogram Max Height (BU)



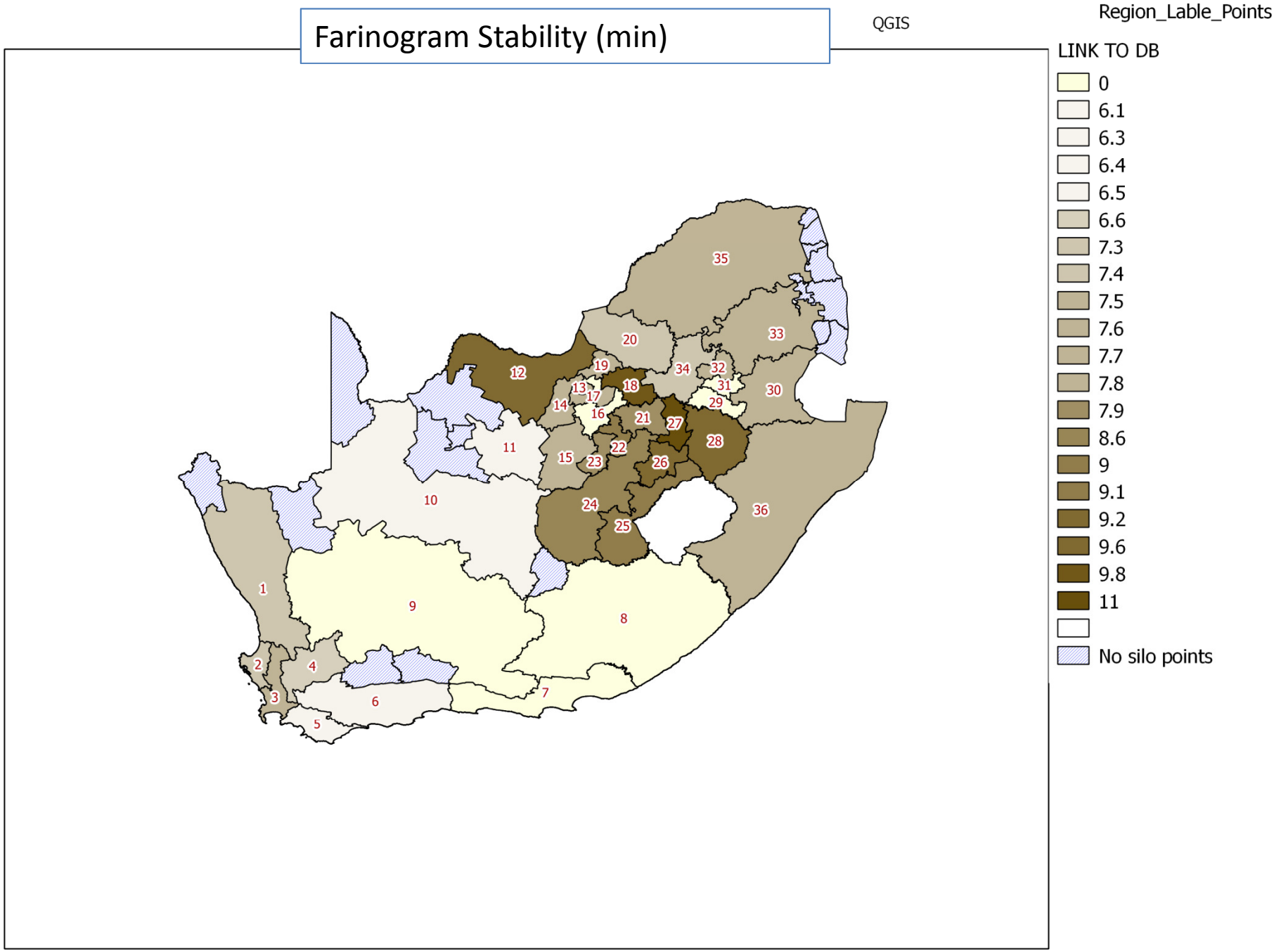
Extensogram Strength (cm^2)



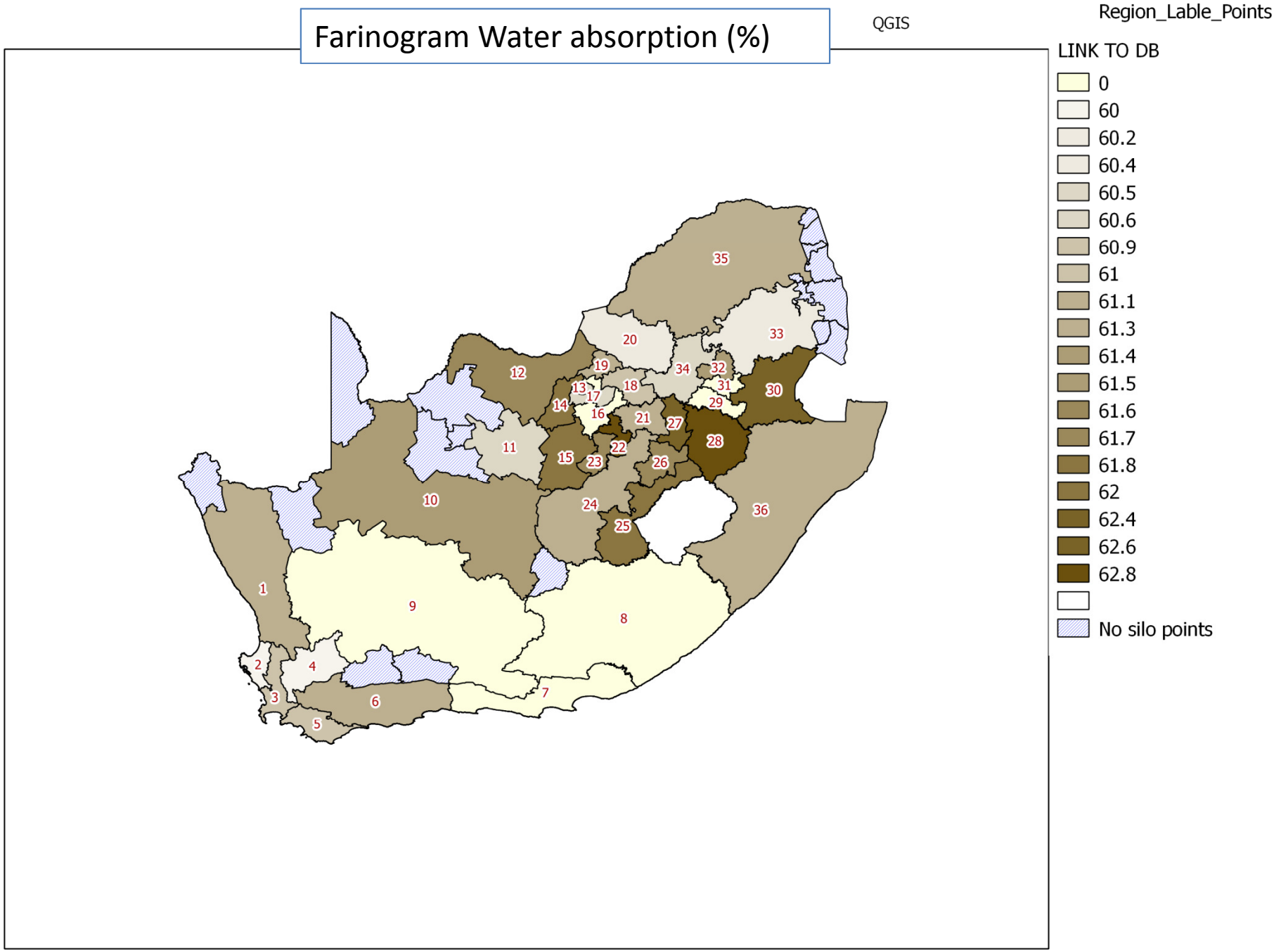
Farinogram Development time (min)



Farinogram Stability (min)



Farinogram Water absorption (%)



CONCLUSION

- ✓ A complete statistical analysis of wheat quality data since the 2003/2004 production season
- ✓ Provided a means to interpret the crop quality analytical data
- ✓ Identify trends to assist with future direction of decisions
- ✓ Goal of the project to present the data in a more accessible fashion has been achieved
- ✓ GIS tool – successfully used for two grain crops namely maize and wheat
- ✓ Big difference between wheat crop quality and maize crop quality is that wheat crop quality is not dependent on season except for a few isolated areas while maize crop quality is heavily dependent on seasonal variations.