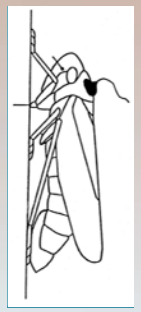
EPG Work-shop



EPG Data Analysis 101

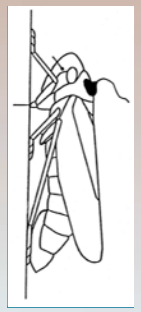
Finding and removing data errors

- by
- T.A. Ebert
- M.E. Rogers

Introduction

- Gathering EPG data involves high levels of concentration on an uninteresting task for long periods of time.
- I cannot think of another set of conditions ideally suited for producing errors.
- We will use a small data file as an example.

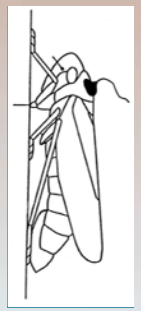
EPG Work-shop



The Strategy

- We will do the following:
  - Open the data file
  - Create an error
  - Save the file
  - Close the file
  - Run the error checking program
  - Examine the result
  - Repeat for other kinds of errors
  - Take a quiz.

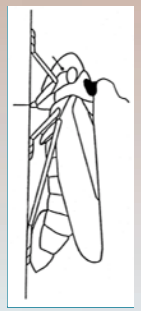
EPG Work-shop



Requirements

- You are familiar with the activities in the Excel skill module.
  - Rapidly moving through data using control key and arrow keys.
  - Selecting large areas
  - The if test in Excel, and nesting if statements
  - Finding the average
  - Global Replace
  - Locking cell references by adding $

EPG Work-shop

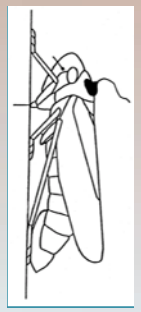


Introducing an Error

- Open the file PsyllidData1.csv
- Insert a new line 6 with
  - Insectnumber = a01
  - Waveform = V
  - Duration = 243
- Run error checker

	A	B	C	D
1	insectno	waveform	Dur	
2	a01	NP	425.3	
3	a01	C	3676.08	
4	a01	NP	1299.84	
5	a01	C	742.4	
6	a01	V	243	
7	a01	G	180.32	
8	a01	C	104	
9	a01	G	511.04	
10	a01	C	57.6	
11	a01	G	144	
12	a01	C	124	
13	a01	NP	1988.64	

EPG Work-shop



A typographical error

- You should see this:

This tells you that there is a waveform V.

It occurs only once.

Here you see that you start with C, go to V, and then go to G

Frequency Table of Waveform Event Transitions

The FREQ Procedure

waveform	Frequency	Percent	Cumulative Frequency	Cumulative Percent
C	234	44.23	234	44.23
D	18	3.40	252	47.64
E1	35	6.62	287	54.25
E2	20	3.78	307	58.03
G	39	7.37	346	65.41
NP	182	34.40	528	99.81
V	1	0.19	529	100.00

Page Break

Frequency Table of Waveform Event Transitions

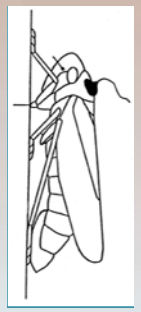
The FREQ Procedure

trans1	Frequency	Percent	Cumulative Frequency	Cumulative Percent
C to D	18	3.47	18	3.47
C to G	38	7.32	56	10.79
C to NP	172	33.14	228	43.93
C to V	1	0.19	229	44.12
D to C	2	0.39	231	44.51
D to E1	16	3.08	247	47.59
E1 to C	15	2.89	262	50.48
E1 to E2	20	3.85	282	54.34
E2 to E1	19	3.66	301	58.00
G to C	39	7.51	340	65.51
NP to C	178	34.30	518	99.81
V to G	1	0.19	519	100.00

Page Break

Duration by waveform Output

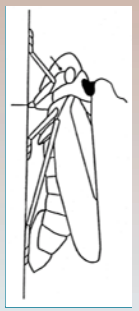
EPG Work-shop



Find the error (approach I)

- Open the data file in Excel
- Select column B
- Ctrl F (for find)
- Enter V
- Press button "Find all"
- Correct problem.
- Save.

EPG Work-shop

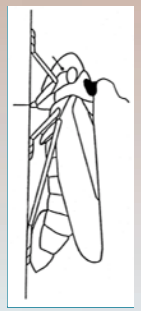


Find the error (approach II)

If you have several of the same error then this approach is more efficient.

- Open the data file in Excel
- Cell D1 enter a 1
- Cell D2 enter =B1+1
- Cell E2 enter =if(b2="V",D2,"")
- Fill down.
- Copy column E2, and paste values back into column E2. Sort.
- The number 6 will be at the top of column E.

EPG Work-shop

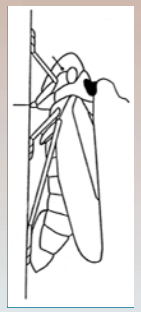


Correcting the problem

- The problem is in row 6.
- If this was a real problem you would have to go back to the original data file and figure out what went wrong.
- In this case, simply delete the row.

- Also delete columns D and E.

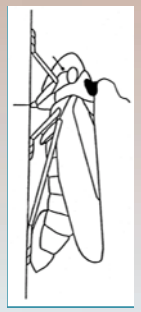
EPG Work-shop



A negative value

- In row 12, put a minus sign in front of the number (-1988.64)
- Save, and close file.
- Run error checker, and get this:

EPG Work-shop



This is the row to examine. It may be one or two rows off.

This is a problem.
Most likely you should simply delete this value. However, you should go back to the EPG recording to make sure.
(In this case, just delete the minus sign.)

Note, that a single error in this case results in two observations appearing as an error.

Duration by waveform Output

Obs	insectno	waveform	dur	w0	w1	in0	marker1
11	a01	NP	-1988.64	C	NP	a01	1
12	a01	C	2977.82	C	C	a01	1

Page Break

Duration by waveform Output

The MEANS Procedure

insectno=a01 waveform=C

Analysis Variable : dur				
N	Minimum	Maximum	Mean	Median
30	2.0800000	6917.30	1113.91	555.9200000

insectno=a01 waveform=D

Analysis Variable : dur				
N	Minimum	Maximum	Mean	Median
3	25.2800000	78.2400000	44.5333333	30.0800000

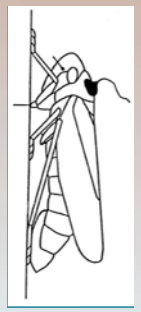
Repeat values

- Insert a new row 11
- Copy line 12, paste into row 11.
- Your file should look like this:
- Save, and close file.

Two C waveforms in a row

	A	B	C	D	E
1	insectno	waveform	Dur		
2	a01	NP	425.3		
3	a01	C	3676.08		
4	a01	NP	1299.84		
5	a01	C	742.4		
6	a01	G	180.32		
7	a01	C	104		
8	a01	G	511.04		
9	a01	C	57.6		
10	a01	G	144		
11	a01	C	124		
12	a01	C	124		
13	a01	NP	1988.64		
14	a01	C	2977.82		
15	a01	NP	336		
16	a01	C	2555.36		

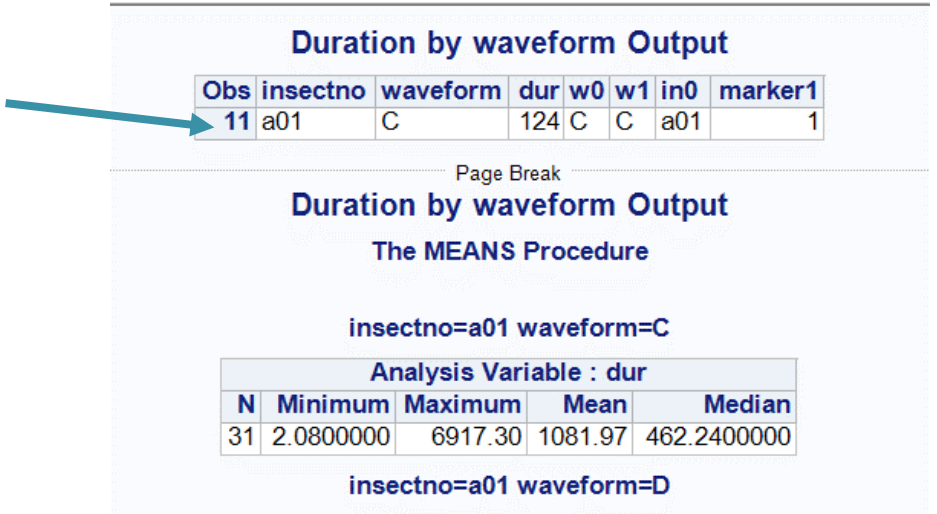
EPG Work-shop



Repeat values

- This is not an obvious error.

Note the observation number.



Duration by waveform Output

Obs	insectno	waveform	dur	w0	w1	in0	marker1
11	a01	C	124	C	C	a01	1

Page Break

Duration by waveform Output

The MEANS Procedure

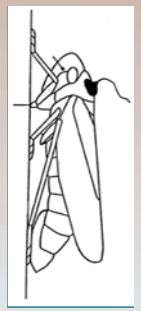
insectno=a01 waveform=C

Analysis Variable : dur				
N	Minimum	Maximum	Mean	Median
31	2.0800000	6917.30	1081.97	462.2400000

insectno=a01 waveform=D

- You just have to go back to the data file and see if there is a problem.
- In this case delete one of the duplicate waveforms.

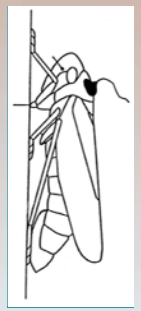
EPG Work-shop



Invalid transitional events

- Transitional events are where the insect changes from one behavior to another.
- Invalid transitional evens are where the insect can't or won't perform a given transition.
- Easy examples are things like Np followed immediately by E2.
- You are the only one that can find these, but Error Checker can help you.

EPG Work-shop

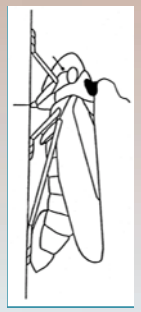


Add an invalid transitional event

- In the data file change the value in B7 from "C" to "E2".
- Your data should now look like this:
- Save, and close the file.
- Run SAS.

	A	B	C	D
1	insectno	waveform	Dur	
2	a01	NP	425.3	
3	a01	C	3676.08	
4	a01	NP	1299.84	
5	a01	C	742.4	
6	a01	G	180.32	
7	a01	E2	104	
8	a01	G	511.04	
9	a01	C	57.6	
10	a01	G	144	

EPG Work-shop



The error

- It is not possible to go directly from G to E2 or from E2 to G.

Frequency Table of Waveform Event Transitions

The FREQ Procedure

waveform	Frequency	Percent	Cumulative Frequency	Cumulative Percent
C	233	44.13	233	44.13
D	18	3.41	251	47.54
E1	35	6.63	286	54.17
E2	21	3.98	307	58.14
G	39	7.39	346	65.53
NP	182	34.47	528	100.00

Page Break

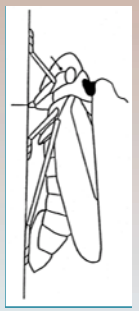
Frequency Table of Waveform Event Transitions

The FREQ Procedure

trans1	Frequency	Percent	Cumulative Frequency	Cumulative Percent
C to D	18	3.47	18	3.47
C to G	38	7.34	56	10.81
C to NP	172	33.20	228	44.02
D to C	2	0.39	230	44.40
D to E1	16	3.09	246	47.49
E1 to C	15	2.90	261	50.39
E1 to E2	20	3.86	281	54.25
E2 to E1	19	3.67	300	57.92
E2 to G	1	0.19	301	58.11
G to C	38	7.34	339	65.44
G to E2	1	0.19	340	65.64
NP to C	178	34.36	518	100.00

Here are the errors.

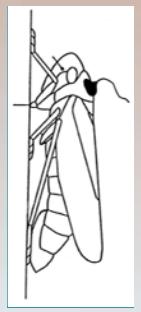
EPG Work-shop



The errors so far

- Typographical error
  - Wrong waveform
- Negative durations
- Repeat values
- Invalid transitional events

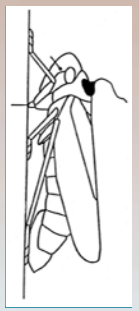
EPG Work-shop



More complex example

- Run errorchecker on the file PsyllidData1A.csv.
- I have added eight errors.
  - Line 46: E1 entered as E2
  - Line 485: 4314.56 changed to negative
  - Line 530: A second E2 was added
  - Line 451: Waveform NP entered as MP
  - Line 190: duplicate of line 191
  - Line 159: NP entered as C
  - Line 65: NP entered as BP
  - Line 65: duration made negative

EPG Work-shop



ErrorChecker output

- Here are the errors.

In this case there are three E2 events in a row. Only the first two are listed here.

Two errors are shown here.

Duration by waveform Output

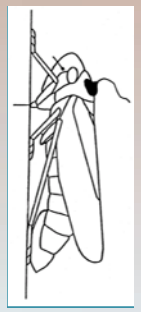
Obs	insectno	waveform	dur	w0	w1	in0	marker1
45	a01	E2	18.24	E2	E2	a01	1
46	a01	E2	47.20	E2	E2	a01	1
64	a01	BP	-279.20	C	BP	a01	1
65	a01	C	927.20	C	C	a01	1
158	a06	C	68.48	C	C	a06	1
159	a06	C	3493.92	C	C	a06	1
190	a06	C	3823.68	C	C	a06	1
485	b06	NP	-4314.56	C	NP	b06	1
486	b06	C	263.20	C	C	b06	1
530	b07	E2	356.29	E2	E2	b07	1

Page Break

Duration by waveform Output

- Go through and correct all of these.
- Save the file and run Errorchecker.

EPG Work-shop



More errors?

- The result should be this.

But there is another error.

Frequency Table of Waveform Event Transitions

The FREQ Procedure

waveform	Frequency	Percent	Cumulative Frequency	Cumulative Percent
C	234	44.32	234	44.32
D	18	3.41	252	47.73
E1	35	6.63	287	54.36
E2	20	3.79	307	58.14
G	39	7.39	346	65.53
MP	1	0.19	347	65.72
NP	181	34.28	528	100.00

Page Break

Frequency Table of Waveform Event Transitions

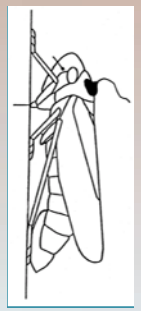
The FREQ Procedure

trans1	Frequency	Percent	Cumulative Frequency	Cumulative Percent
C to D	18	3.47	18	3.47
C to G	39	7.53	57	11.00
C to MP	1	0.19	58	11.20
C to NP	171	33.01	229	44.21
D to C	2	0.39	231	44.59
D to E1	16	3.09	247	47.68
E1 to C	15	2.90	262	50.58
E1 to E2	20	3.86	282	54.44
E2 to E1	19	3.67	301	58.11
G to C	39	7.53	340	65.64
MP to C	1	0.19	341	65.83
NP to C	177	34.17	518	100.00

Page Break

Duration by waveform Output

EPG Work-shop



Success?

- You should now have this result.

Frequency Table of Waveform Event Transitions

The FREQ Procedure

waveform	Frequency	Percent	Cumulative Frequency	Cumulative Percent
C	234	44.32	234	44.32
D	18	3.41	252	47.73
E1	35	6.63	287	54.36
E2	20	3.79	307	58.14
G	39	7.39	346	65.53
NP	182	34.47	528	100.00

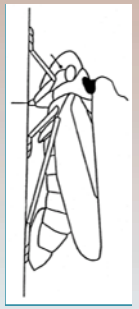
Page Break

Frequency Table of Waveform Event Transitions

The FREQ Procedure

trans1	Frequency	Percent	Cumulative Frequency	Cumulative Percent
C to D	18	3.47	18	3.47
C to G	39	7.53	57	11.00
C to NP	172	33.20	229	44.21
D to C	2	0.39	231	44.59
D to E1	16	3.09	247	47.68
E1 to C	15	2.90	262	50.58
E1 to E2	20	3.86	282	54.44
E2 to E1	19	3.67	301	58.11
G to C	39	7.53	340	65.64
NP to C	178	34.36	518	100.00

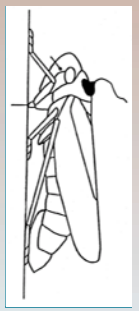
EPG Work-shop



NP start error

- One code must indicate the non-probing behavior.
- This must be the first waveform in every recording.
- Open PsyllidData1.csv
- Insert a new row 2.
- Copy what is now row 4 into row 2.
- Save, close file, run Error Checker.

EPG Work-shop



Note, Error Checker programs downloaded before 10-25-2016 will not catch this problem.

NP not first results

The modified data file looks like this.

	A	B	C	D
1	insectno	waveform	Dur	
2	a01	C	3676.08	
3	a01	NP	425.3	
4	a01	C	3676.08	
5	a01	NP	1299.84	
6	a01	C	742.4	
7	a01	G	180.32	
8	a01	C	104	

The ErrorChecker output looks like this.

Duration by waveform Output

Obs	insectno	waveform	dur	marker1	w0	w1	in0	marker2
1	a01	C	3676.08	1		C	a01	0

Page Break

Duration by waveform Output

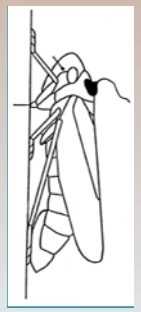
The MEANS Procedure

insectno=a01 waveform=C

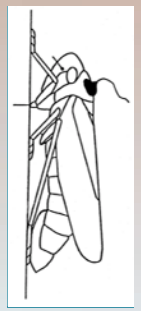
Analysis Variable : dur				
N	Minimum	Maximum	Mean	Median
31	2.0800000	6917.30	1196.56	649.6000000

insectno=a01 waveform=D

EPG Work-shop



EPG Work-shop



Warning: Hidden Error

- This applies to both Backus 2.0 and Ebert.
- **Only one Non-probing behavior is allowed.**
- Multiple non-probing behaviors will cause problems because neither program will correctly identify a probe.
- **Solution**: Analyze non-probing events first. Then combine all non-probing events into a single waveform and rerun the analysis.

Quiz Time

Use Excel and SAS to complete the quiz.

- Analyze file AphidData1err.

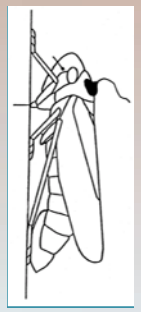
This is what you should have once there are no more errors in file AphidData1err.

Using ErrorChecker and the tools gained in the Excel Skill Module, answer the following question:

What is the total duration of pd for each insect?

Before you start, read the next two slides.

EPG Work-shop



Frequency Table of Waveform Event Transitions

The FREQ Procedure

waveform	Frequency	Percent	Cumulative Frequency	Cumulative Percent
C	1120	49.62	1120	49.62
E1	20	0.89	1140	50.51
E2	9	0.40	1149	50.91
F	6	0.27	1155	51.17
G	2	0.09	1157	51.26
NP	95	4.21	1252	55.47
PD	1005	44.53	2257	100.00

Page Break

Frequency Table of Waveform Event Transitions

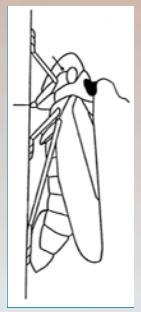
The FREQ Procedure

trans1	Frequency	Percent	Cumulative Frequency	Cumulative Percent
C to E1	18	0.80	18	0.80
C to F	6	0.27	24	1.07
C to G	2	0.09	26	1.16
C to NP	87	3.87	113	5.02
C to PD	1005	44.69	1118	49.71
E1 to C	11	0.49	1129	50.20
E1 to E2	9	0.40	1138	50.60
E2 to C	5	0.22	1143	50.82
E2 to E1	2	0.09	1145	50.91
F to C	6	0.27	1151	51.18
G to C	2	0.09	1153	51.27
NP to C	93	4.14	1246	55.40
PD to C	1002	44.55	2248	99.96
PD to NP	1	0.04	2249	100.00

Some Rules I

- This is an exercise. Ideally, errors are corrected by going back to the original recordings and making sure that the data match the recording.
- Different insects have different waveforms, and different rules.
- These rules are just to work with this data file and example.

EPG Work-shop



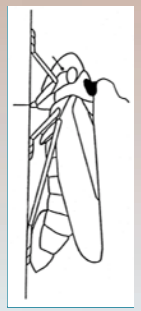
Some Rules II

- Negative values: remove the negative sign.
- The same waveform on multiple lines should have the durations added, and one line deleted.

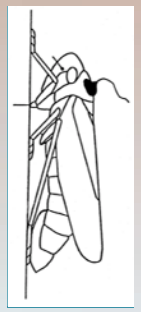
A01	C	452	Should be	A01	C	575
A01	C	123				

- Typographic errors
  - Pd may be pf, ps, pe
  - E2 can be E3, W2, R3
  - Np can be NO, NI, MP
- Only E1 can come directly before E2. Any other result before E1 should be changed back to E1.
- If two identical waveforms have the same duration, then delete one row.

EPG Work-shop



EPG Work-shop



Quiz Part 2

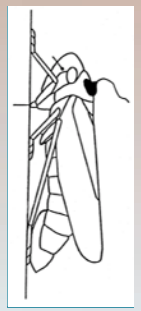
(Note: Carefully read this slide and the next slide before starting.)

- Make a copy of the file MysteryData1.csv
- Open the copy in Excel.
- Run error checker on the original.
- Paste results into the copy.
- Make changes to the original.
- Make sure that you save and close the original Excel file.
- Record the insect numbers where changes were made.
- There are 20 changes, so the correct answer will consist of 20 insect numbers.
- You must follow the rules on the next slide.

Rules for Quiz Part 2

- The only valid waveforms are: NP, C, D, E1, E2, and G
- D, E1, and E2 cannot go directly to G.
- G cannot be right before D, E1, or E2.
- D must come right before E1.
- E1 must come right before E2.
- C must come right before D.
- Only C can follow NP.
- Only one change is needed to fix any mistake.
- If there are two consecutive durations that are the same, delete one row to remove the duplicate.
- HINT: Use Excel to find errors if SAS does not tell you where the error is located.

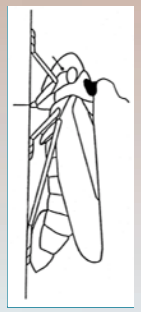
EPG Work-shop



Final Suggestions I

- There comes a point where it is faster to skip SAS and simply work in Excel.
  - If you have a data file like MysteryData1
  - If there are dozens of errors (or hundreds), then it is faster to use Excel to correct most of them.
    - Use find or find and replace.
    - Use if statements to search for problems.
    - Use SAS error checker to find different sorts of problems. Error checker finds 40 negative values? Use Excel to skip through all instances of a negative value.
    - Using the sort feature in Excel might help, but you will loose the contextual information that could help you figure out the correct solution to the problem. It is possible to make a correction that results in another mistake, often one that is harder to find.
  - Use SAS error checker as the final step to make sure that nothing was missed.

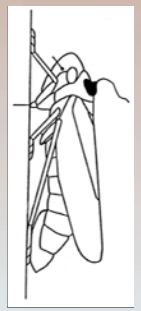
EPG Work-shop



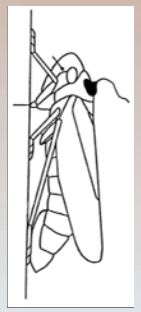
Final Suggestions 2

- Some mistakes are impossible to catch.
  - If some data are entered by hand, then 799.84 could become 7998.4 or 798.94, or some other value.
  - Any data entered by hand should be checked carefully.
  - Sometimes this sort of mistake could manifest as outliers in data analysis.
  - This problem should be very unusual considering that most EPG data entry is handled by the computer.

EPG Work-shop



EPG Work-shop



Possible answers to questions

If you have the right answer then your answer will perfectly match one of the following four columns.

a1	1082.95
a2	338.24
a3	989.82
b1	253.35
b2	783.61
b3	253.35
b4	1082.95
b5	783.61

a1	989.82
a2	1082.95
a3	783.61
b1	253.35
b2	1082.95
b3	783.61
b4	338.24
b5	810.84

a1	1082.95
a2	253.35
a3	989.82
b1	783.61
b2	338.24
b3	989.82
b4	253.35
b5	1082.95

a1	989.82
a2	1082.95
a3	338.24
b1	1082.95
b2	253.35
b3	989.82
b4	783.61
b5	1082.95

a1	a1	a1	a1
a2	a1	a1	a1
a3	a3	a3	a3
a4	a4	a4	a4
b6	b6	b6	b6
c12	c12	c14	c12
c15	c15	c15	c15
d19	d19	d19	d19
e2	e2	e2	e2
f12	f12	f12	f12
f12	f12	f12	f12
g5	g5	g5	g5
h21	h21	h21	h21
k107	k107	k107	k107
L6	L6	L6	L5
m9	m9	m9	m9
n16	n16	n16	n16
n2	n2	n2	n2
o4	o4	o4	o4
O9	O9	O9	O9