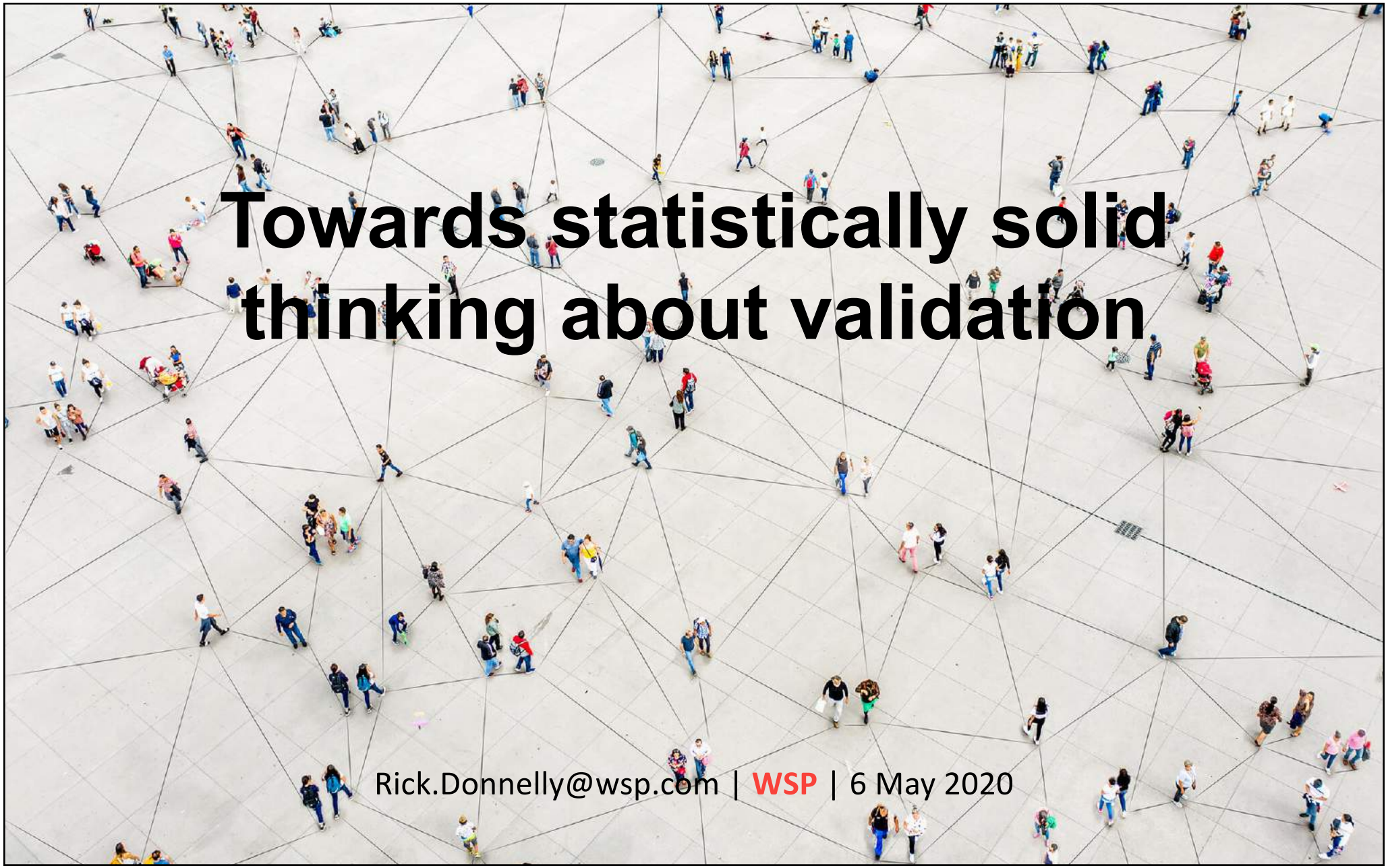


Towards statistically solid thinking about validation

Rick.Donnelly@wsp.com | WSP | 6 May 2020

Traditional practice

		Traditional trip-based models												*Advanced*			
Category	Criteria	AL	AZ	CA	FL	IN	LA	MS	MD	NC	TN	TX	WI	OH	OR	n	Rank
Trip generation	Aggregate trip rates				X								X			2	9
	Trips by purpose				X				X	X		X	X		X	6	5
Trip distribution	Travel time		X										X		X	3	8
	Average trip lengths				X				X	X		X	X		X	6	5
Mode choice	Auto occupancy			X	X		X	X	X	X		X			X	8	4
	Mode share			X					X	X					X	4	7
Traffic assignment	Area type (volume vs. count)		X	X	X	X					X	X				6	5
	Corridor (volume vs. count)					X				X				X		3	8
	District/region (volume vs. count)			X	X		X	X		X	X	X	X			8	4
	Facility type (volume vs. count)	X	X	X	X	X	X	X			X	X	X			10	2
	Intercity transit ridership								X				X		X	3	8
	GEH statistic												X			1	10
	RMSE by volume group		X		X	X			X	X	X		X	X	X	9	3
	R-squared (regression)		X													1	10
	Screenline (volume vs. count)		X		X	X	X	X		X	X	X	X	X		10	2
	Tonnage/commodity flows				X					X		X	X		X	5	6
	Total volume				X								X			2	9
	Trucks (volume vs. count)				X				X				X			3	8
	VHT (volume vs. count)				X										X	2	9
	VMT (volume vs. count)		X		X	X	X	X	X	X	X	X	X	X	X	12	1
	Volume group (volume vs. count)	X			X	X					X		X	X		6	5

Source: NCHRP 08-36B, Task 91, "Validation and sensitivity considerations for statewide models" (2011); data from shaded states added by Rick Donnelly on 30-Jan-2016

Top five criteria (n=9)

Rank	Criteria	Category
1	VMT (volume vs. count)	Traffic assignment
2	Observed vs. simulated by facility type	Traffic assignment
	Observed vs. simulated by screen line	Traffic assignment
3	(P)RMSE by volume group	Traffic assignment
4	Observed vs. simulated by district/region	Traffic assignment
	Auto occupancy	Pre-assignment
5	Observed vs. simulated by volume group	Traffic assignment
	Total trips by person	Trip generation
	Average trip length by trip purpose	Destination choice

Three revelations

Rethinking validation:

1. Differentiate by count type
2. Trimming outliers
3. Using meta-networks

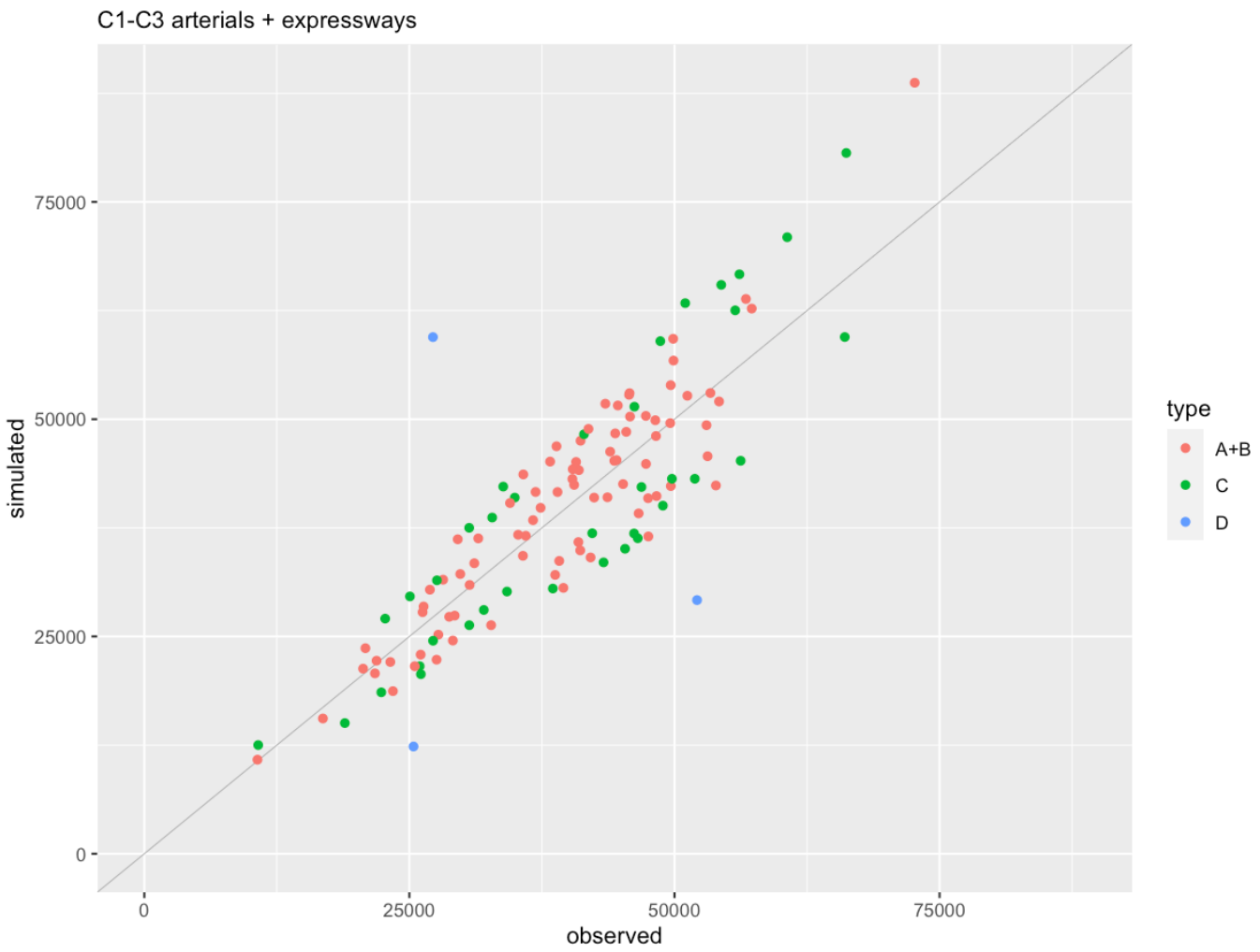
Not all counts are created equal

Type	Description	Weight
A	Permanent traffic recorders	4.0
B	Temporary piezoelectric or vehicle classifiers	2.0
C	24+ hour pneumatic tube counts	1.0
D	Everything else (manual counts, peak periods factored to daily, …)	0.1 - 0.5

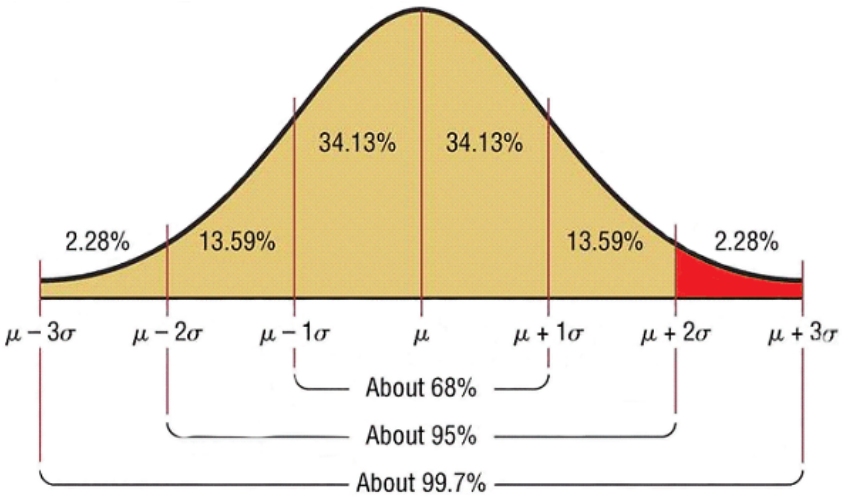
Strategies:

- Jettison type D counts if possible
- Multiple summaries: A+B only, A-C only, all
- Only use A+B or A-C to assess model performance
- Look at A-D (all counts) patterns to spot patterns
- Type D counts never disqualifies a model

Recent outcome



Potential error in counts



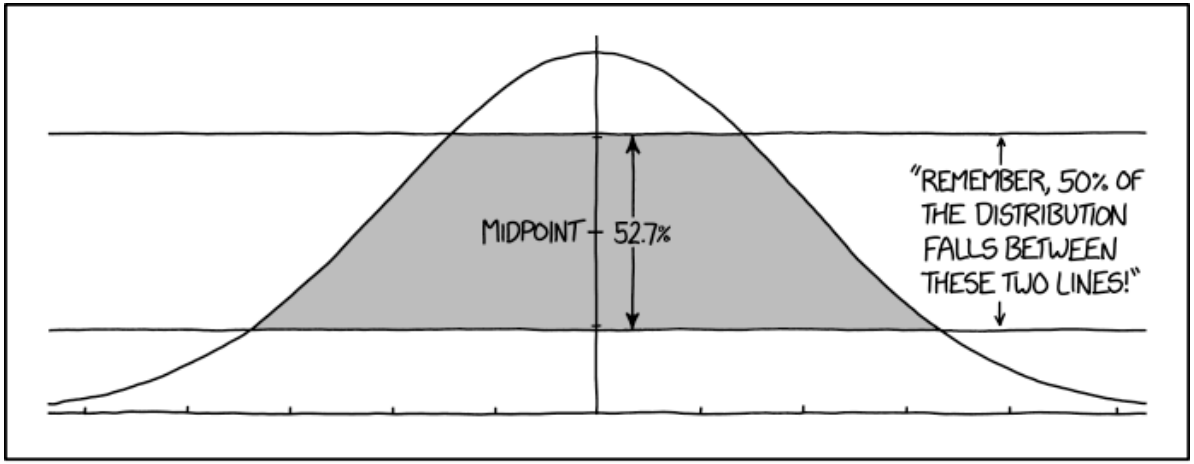
Implications for counts:

Obs	±1σ	±2σ	±3σ
40	27	11	2
80	55	22	4
120	82	33	5
160	109	43	7
200	137	54	9
240	164	65	11

Assumptions:

- Data are normally distribution
- Distribution has mean μ and standard deviation σ
- Observations are independent of one another

Alternative view



HOW TO ANNOY A STATISTICIAN

Some traditional stats

Root mean squared error:

$$RMSE = \sqrt{\frac{\sum_k (O_k - S_k)^2}{n-1}}$$

Percent root mean squared error:

$$PRMSE = \left(RMSE \div \overline{O}_k\right) \times 100.0$$

Outlier effects on RMSE and percent RMSE

(a) Illustrative data

observed	simulated	difference	difference^2
12910	13505	595	354025
10821	9932	-889	790321
16494	13873	-2621	6869641
11111	10431	-680	462400
12650	14001	1351	1825201
13805	12644	-1161	1347921
9999	11748	1749	3059001
10000	9200	-800	640000
10500	10900	400	160000
	Total (sum of squared error)		15508510
	Number of observations		9
	Root mean squared error (RMSE)		1392.3
	Mean of observed values		12032.2
	Percent RMSE		11.6

Standard deviations:

Measure	Value
1 SD	2154.3
2 SD	4308.7
3 SD	6463
4 SD	8617.4

(b) Replace one value with outlier

observed	simulated	difference	difference^2
12910	13505	595	354025
10821	9932	-889	790321
16494	13873	-2621	6869641
11111	10431	-680	462400
35003	14001	-21002	441084004
13805	12644	-1161	1347921
9999	11748	1749	3059001
10000	9200	-800	640000
10500	10900	400	160000
	Total (sum of squared error)		454767313
	Number of observations		9
	Root mean squared error (RMSE)		7539.6
	Mean of observed values		14515.9
	Percent RMSE		51.9

Outlier position:

Observed / simulated	2.5
Observed - simulated	21002
Standard deviations	9.7

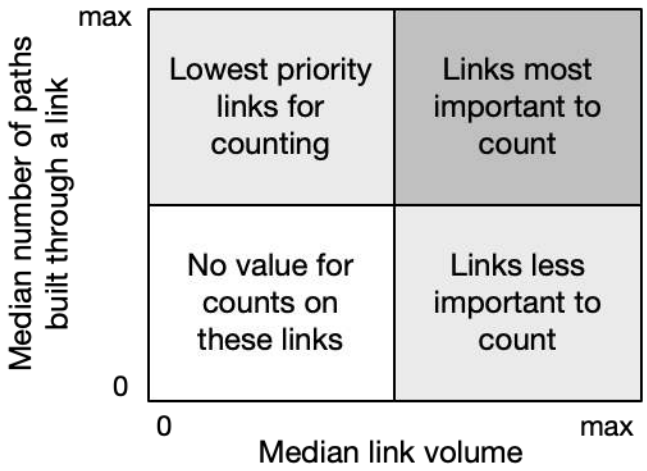
(c) Trimmed RMSE case

observed	simulated	difference	difference^2
12910	13505	595	354025
10821	9932	-889	790321
16494	13873	-2621	6869641
11111	10431	-680	462400
~~12650~~	~~31625~~	0	0
13805	12644	-1161	1347921
9999	11748	1749	3059001
10000	9200	-800	640000
10500	10900	400	160000
	Total (sum of squared error)		13683309
	Number of observations		8
	Root mean squared error (RMSE)		1398.1
	Mean of observed values		12032.2
	Percent RMSE		11.6

Optimal counts?

Network design process for identifying optimal truck count locations

y-axis: *A scalar matrix (1.0) is assigned, with the resulting link flows being the number of OD paths passing over it; links with larger number of paths than the median are more important to count.*



These links have highest assigned truck flows and highest number of paths built through them, so are most useful in model development and validation

x-axis: *The seed matrix (estimated truck OD flows) is assigned to network; links with flows above the median assigned link flows are more important to count than those below the median.*

Meta-networks

1. Include only freeways, expressways, major arterials
2. Collapse freeway interchanges to single point
3. Meta-nodes: where these arcs intersect
4. Meta-links: single bidirectional arc between meta-nodes
5. One count per meta-link → independent observations by definition
6. Validate and visualize at the meta-network level

Forecasting errors

Unexpected changes in exogenous information
Socioeconomic forecasts
Vehicle technology and efficiency
Land use
…

Political influences
Plans and policies
Taxation
Regulations
Mismatch between preferences and actions

Methodological
Patterns change faster than models
System complexity
Aggregation bias
Discontinuities
Under-specified behaviour
Over-fitting models

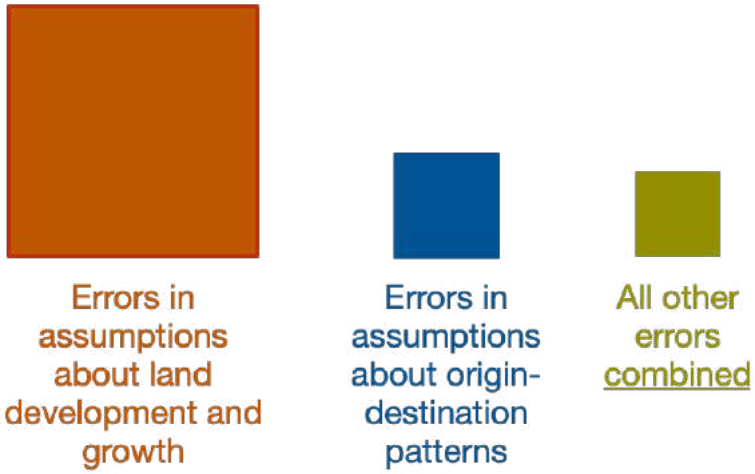
Irrational behaviour
Optimism bias
Agency mission conflicts

Poor data
Errors (specification, sampling, …)
Cross-sectional
Inferential
RP-SP disconnect
Unwillingness or inability to invest
Institutional barriers to sharing

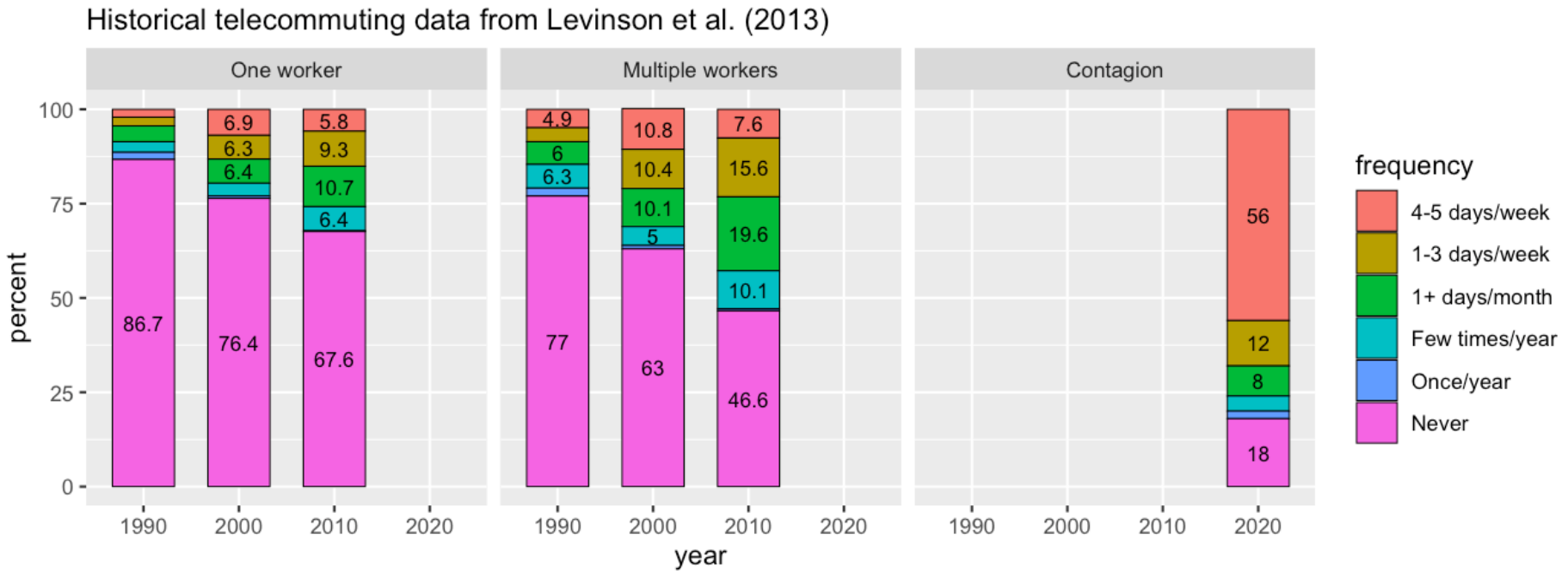
Deception
Weak accountability
Bad incentives
"Delusion as necessity"
No code of ethics
Lack of consequences for engaging in deception

Adapted from Flyvbjerg et al (2003). Italicized entries added by R Donnelly.

In perspective...



Telecommuting trends over time



Questions?