GPS-Paradata in Computer-Assisted Personal Interviews: Additional Opportunities for Monitoring Fieldwork Interviewers

8th European Survey Research Association conference
15-19 July, 2019
Zagreb, Croatia

Daniil Lebedev,
Department of Sociology, National Research University
"Higher School of Economics"

Background

Separate locations analysis:

- "Geofencing" - Comparison of locations at the beginning and at the end of an interview [Seeger 2011; Wang, Biemer 2010; Mohajer, Edwards 2018; Choumert-Nkolo et al. 2019]

- "Strand length" - Comparison of interview location and that of the sampled houshould [Mohajer, Edwards 2018; Sikes 2009] or with interviewers own home [Hasson 2015]

- "Curbstoning" test - checking for presence of too dense groups of interviews' locations [U.S. Census 2010; Dajani, Marquette 2015]

Route analysis (sequence of GPS measured locations):

- Linking locations of interviews into the route [Choumert-Nkolo et al. 2019]

- Analysis of interviewers' routes [Wagner, Olson, Edgar 2017; Olson, Wagner 2015]

Research question:

How to use GPS-paradata in computer-assisted personal interviews on tablets for fieldwork monitoring?

Data

26th wave, RLMS HSE, CAPI, 37 interviewers, **7** regions, **491** interviews

October 2017– February 2018

Individual surveys

GPS locations – information regarding latitude and longitude of a tablet in the beginning and at the end of the interview (SurveySolutions application) – active measurement

GPS routes – information about interviewers' routes in the field (GPSLogger application) – passive measurement

Additional data:

- Interviewers' socio-demographic characteristics and expectations towards the success of transition

GPS-paradata quality

Missing data

Measurement accuracy (GPS Logger, Survey Solutions)

Missing data

Binary logistic regression

Dependent variable: Missing data of location measurements either at the beginning or at the end of the interview - 110 cases (22,4%)

	B	S.E.	Sig.	Exp(B)	B	S.E.	Sig.	Exp(B)
Age	0,87	0,40	**0,031**	2,39	4,41	2,39	0,065	82,51
Tablet availability	1,26	0,38	**0,001**	3,52	0,95	0,74	0,202	2,58
Confidence with tablet	-0,48	0,28	0,086	0,62	-2,69	1,15	**0,020**	0,07
Expectations index	-1,53	0,37	**0,000**	0,22	-3,42	1,85	0,065	0,03
Mean accuracy (GPS Logger)	-0,75	0,28	**0,006**	0,47	2,79	1,27	**0,028**	16,20
Mean battery charge level	-0,87	0,28	**0,002**	0,42	2,30	1,39	0,098	9,92
Solikams					-13,86	5,69	**0,015**	0,00
Kazan					-22,30	4893,87	0,996	0,00
Kurgan					-4,92	3,88	0,205	0,01
Volsk					6,68	5,33	0,210	797,16
Moscow region					-9,52	4,37	**0,030**	0,00
Berdsk					2,63	2,34	0,262	13,80
Constant	-2,66	0,35	0,000	0,07	-0,77	1,61	0,632	0,46
-2 Log likelihood		127,97				106,101		
Cox & Snell R Square		0,224				0,283		
Nagelkerke R Square		0,438				0,553		

GPS-paradata accuracy

GPS Logger	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B		Survey solutions	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound		B	Std. Error	Beta			Lower Bound	Upper Bound
(Constant)	-0,66	0,09		-7,18	0,000	-0,84	-0,48	**(Constant)**	24,61	1,4		17,58	0,000	21,86	27,36
Solikamsk	1,48	0,15	0,48	10,09	**0,000**	1,20	1,77	**Solikamsk**	-0,24	2,24	-0,01	-0,11	0,914	-4,64	4,16
Kazan	1,47	0,16	0,43	9,32	**0,000**	1,16	1,79	**Kazan**	-3,21	2,39	-0,08	-1,34	0,180	-7,91	1,49
Kurgan	1,12	0,13	0,43	8,56	**0,000**	0,86	1,38	**Kurgan**	-1,17	1,99	-0,04	-0,59	0,558	-5,07	2,74
Volsk	-0,68	0,16	-0,20	-4,28	**0,000**	-1,00	-0,37	**Volsk**	-0,73	2,52	-0,02	-0,29	0,773	-5,68	4,22
Berdsk	0,40	0,15	0,13	2,69	**0,007**	0,11	0,69	**Berdsk**	0,50	2,31	0,01	0,22	0,828	-4,04	5,04
Moscow region	0,74	0,12	0,31	6,00	**0,000**	0,50	0,98	**Moscow region**	-1,16	1,81	-0,05	-0,64	0,520	-4,71	2,39

	Mean accuracy (GPSLogger)	Mean accuracy (Survey solutions)
R	0,654	0,084
R Square	0,428	0,007
Adjusted R Square	0,419	-0,009
Std. Error of the Estimate	0,763	11,457

Identifying "suspicious"/ "at risk" interviews

Comparing location at the beginning and at the end of the interview

Thresholds choice

Time difference in location measurements

Geofencing

Two ways of threshold identification:

Conventional (8-25 metres) [Keating et al. 2014] – distance is more than 50 metres

48 (12,6%)   **35 (9,2%)**   **45 (11,8%)**
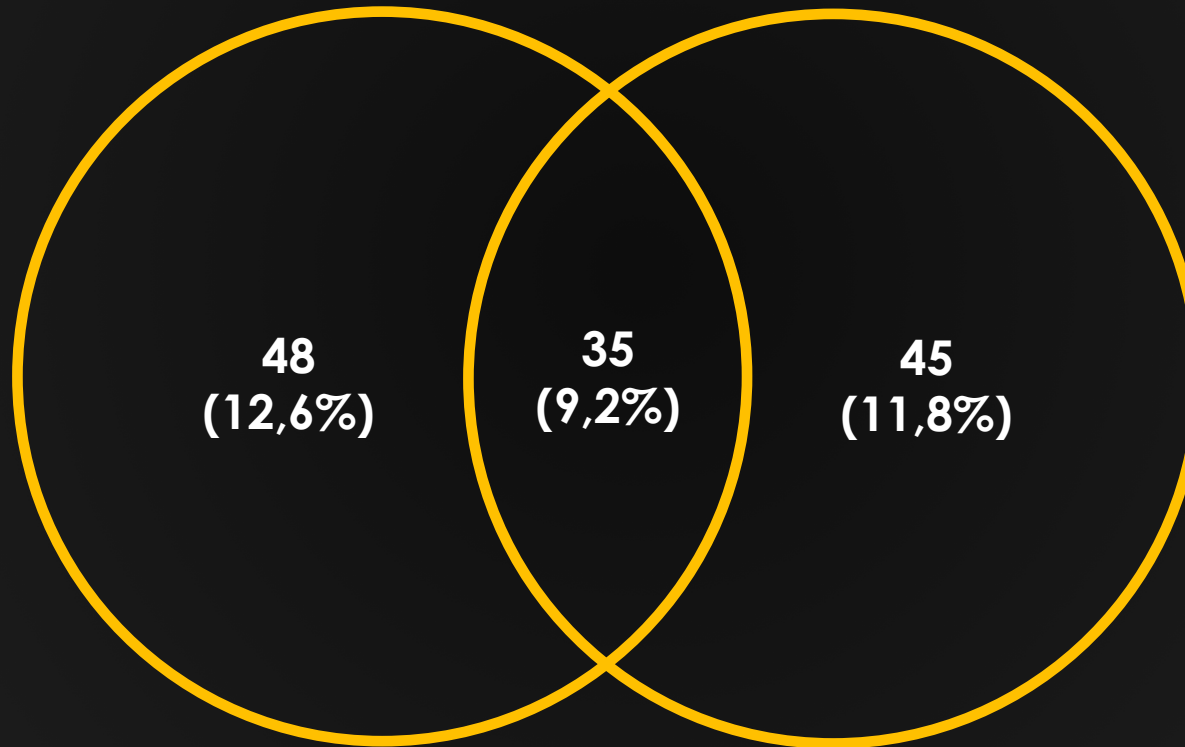
Accuracy-based (Survey Solutions) – distance is more than sum of accuracy of location measurements

Overall "suspicious" interviews (distance based) – **58 (15,2%)**

Significant distance between two locations

Binary logistic regression
Dependent variable: Significant distance between two locations (either methods of thresholds identification)

	B	S.E.	Sig.	Exp(B)
Satisfaction with RLMS project	0,63	0,37	0,086	1,88
Tablet	0,29	0,37	0,424	1,34
Confidence with tablet	-0,62	0,28	**0,024**	0,54
Expectations index	-0,06	0,26	0,822	0,94
CAPI experience	0,66	0,27	**0,014**	1,93
Age	-0,53	0,29	0,063	0,59
Mean satellites number (GPS Logger)	-0,05	0,23	0,846	0,96
Mean time difference between locations (GPS Logger)	-0,19	0,25	0,431	0,82
Mean battery charge level	0,09	0,27	0,748	1,09
Constant	-1,68	0,24	0,000	0,19
-2 Log likelihood	200,863			
Cox & Snell R Square	0,067			
Nagelkerke R Square	0,113			

Negative time difference between location measurements

Negative time difference between time of location measurements at the beginning and at the end of interview

49 (12,9%)   **11 (2,9%)**   **58 (15,2%)**
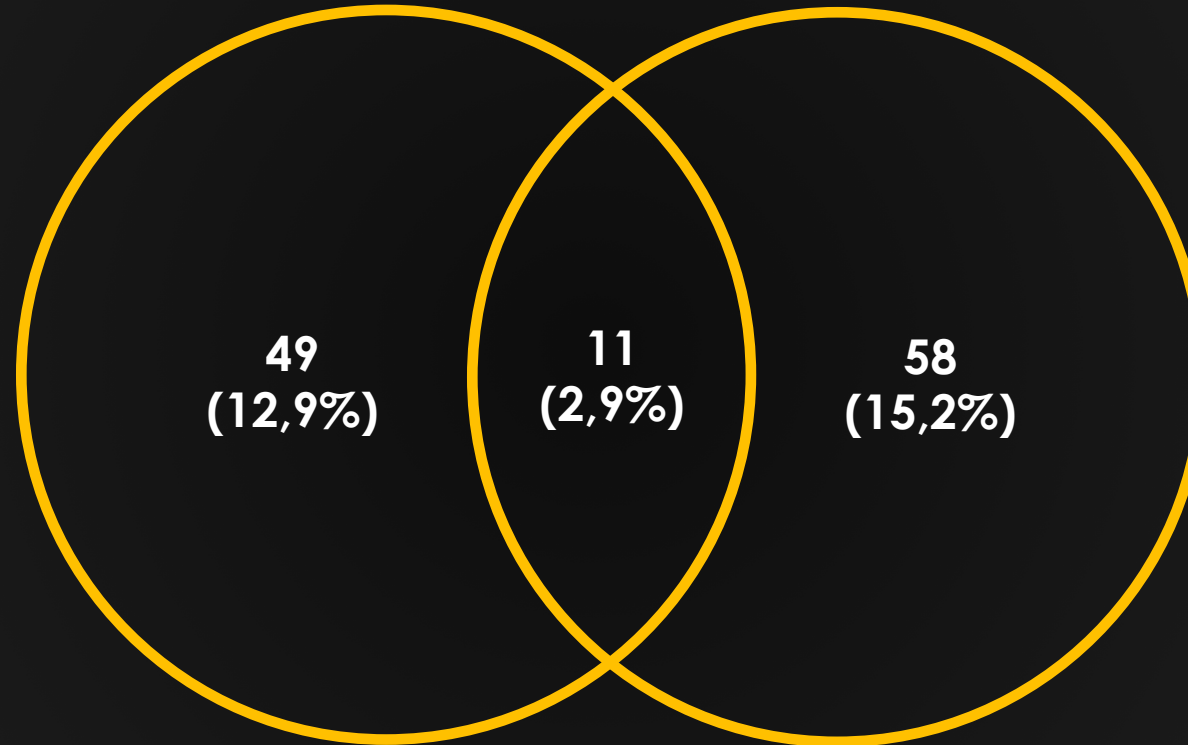
Significant distance between locations (either methods of threshold identification)

Overall "suspicious" interviews – **96 (25,2%)**

Curbstoning analysis

Distance between interviews conducted in different households:

	Frequency	Percent
Distance between interviews is less than 8*	90	18,3
Distance between interviews is less than 16*	132	26,9
Same HH interviews have distance of more than 16	25	5,1
Missing data	94	19,1
No other members from this household were interveiwed	85	17,3
* - (in more than 2 cases)		

Outlook (1)

Four indicators of «suspicious» interviews:

- Significant distance between locations at the beginning and at the end of an interview (conventional threshold identification (48 – 12,6%) and accuracy-based (45 – 11,8%))

- Negative time difference between location measurements (49 – 12,9%)

- Interviews proximity (excluding members of the same HH) by 8 metres (90 – 18,3%) or by 16 metres (132 – 26,9%)

- Significant distance between interviews within the same HH (25 – 5,1%)

Overall «suspicious»/«at risk» interviews – (274 – 55,8%)

Outlook (2)

GPS-paradata quality may vary in connection with regions (lower quality in more developed regions [Lemmens 2011; Gong et al. 2012]) and with interviewers' characteristics (confidence with CAPI)

Experience with CAPI is connected with higher probability of significant distance between locations at the beginning and at the end of the interview (by 93%), while high levels of confidentiality regarding working with tablet is connected with lower probability of existence of significant distance (by 46%)

Recommendations

Focus on interviewers education while starting using CAPI which can be connected with further increase in GPS-paradata quality and with lower levels of «suspicious» interviews

Use accuracy as threshold identification for distance between two loactions (e.g. at the beginning and ant the end of the interview) – GPS-data quality may vary in different regions

GPS-paradata should be used in conjunction with other methods of fieldwork monitoring – no exact assumptions about fabrications or falsifications may be done based on GPS-paradata analysis only (nonintentional errors or technical difficulties)

Limitations

GPS paradata employment differs in case of longitudinal panel and cross-sectional surveys as well as between surveys with different sampling design

We were unable to use respondents' addresses

Regarding passive GPS-data capturing we were unable to detect whether some additional software was used and when it was turned off

Future plans

Second wave of experimental RLMS-HSE CAPI – additional regions, interviewers, respondents and data

Location comparison between waves (panel option)

Compare applications (GPS Logger, Survey Solutions) from the standpoint of data quality, precision

Thank you for the attention

For cooperation, questions and comments please contact:
zenon-daniil@yandex.ru

Thnanks to Dr. Aigul Mavletova for considerable impact on this research

Contents