Liv Tollånes
Stage, FOCOS, spring 2024

Project on the SES Embedding Method: Figures, Tables, and Outputs

Model	Markers	Followers (n)
m1	All	115,079
m2	Remove political party	115,048
m3	Remove political party, media, university	105,582
m4	Remove media, political party, university, football clubs	101,885
m5	Remove media, political party, university, luxury clothing and malls, and football clubs	100,650
m6	Remove media, party, university, clearly luxury brands, and football	100,869
m7	Include only things you can buy (chain restaurants, commercial retail, supermarket chains, luxury clothing and malls, sportswear)	51,322
m8	Include magazine and media only	108,843
m9	Education related markers only (universities, business schools, and professional high schools)	36,577

Table 1: Overview of models used to assess the optimal combination of markers for SES inference.

	Marker	Type	In-degree centrality
1	Mediapart	Information	55.7%
2	BFMTV	Information	45.7%
3	LEXPRESS	Information	35.4%
4	Franceculture	Information	34.8%
5	Europe1	Information	34.2%

Table 2: Markers ranked by their in-degree centrality, indicating the percentage of incoming edges in the full network directed towards each node.

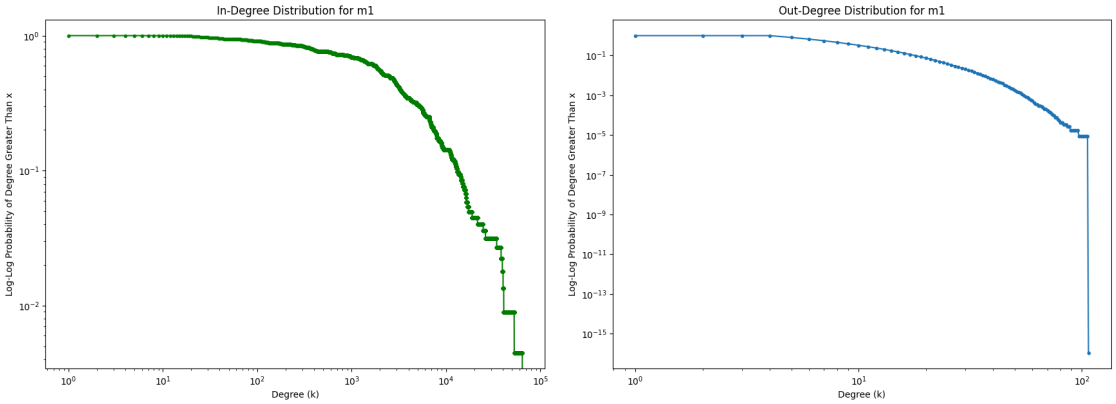


Figure 3: Complementary Cumulative Distribution Functions (CCDF) of in-degrees and out-degrees for markers and followers in the full bipartite graph (m1). The in-degree represents the number of followers that markers have, indicating their popularity. The out-degree represents the number of markers that users follow. The plots are displayed on a log-log scale. The probabilities decrease as the in-degrees and out-degrees increase, with a steeper decline observed for markers. This steep drop suggests the presence of a few highly influential markers, as they have significantly higher in-degrees compared to others.

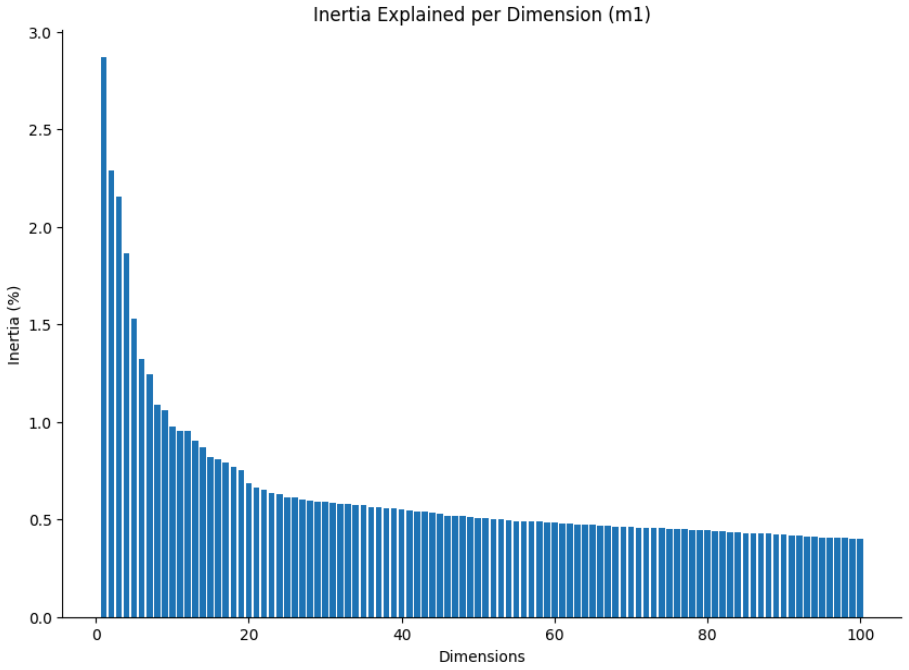


Figure 4: Percentage of inertia explained per CA dimension for the full model, m1. Most of the explained inertia is captured in the first dimensions.

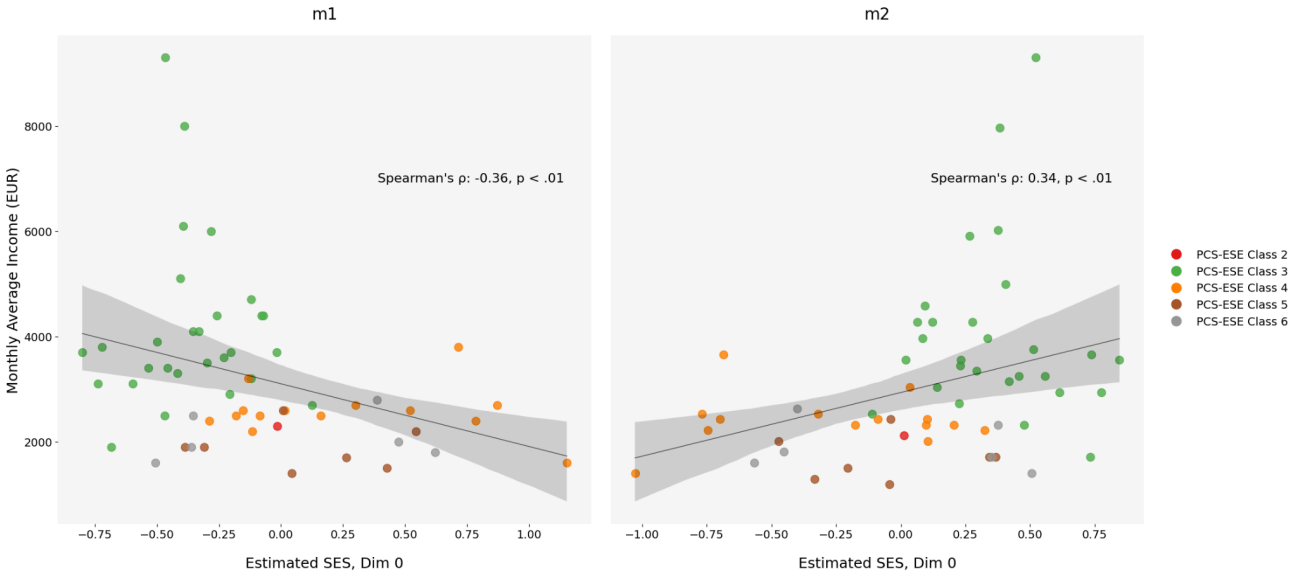


Figure 5: Relationship between median SES estimates per PCS-ESE group and their associated income for users in the validation set. Only models 1 and 2 showed a significant relationship (p<.05). Points are coloured according to the main groups of the PCS-ESE scale, with each point representing the median SES estimate for a specific PCS-ESE group in the validation data. Due to the potential rotation of axes during CA, the SES estimates must be interpreted contextually. The numerical values on the x-axis, whether positive or negative, do not inherently represent high or low SES. The mirroring effect between m1 and m2 indicates a common underlying structure in the first dimension. The y-axis shows the average full-time equivalent salary in EUR of PCS-ESE groups per month.

Model nr.	***Mean RMSE (CV***	***R2 (CV)***	***R2 (Full)***	***Max Coeff Predictor***	***Max Coeff Value***
1	1438.970	0.877	0.997	0	-153.529
2	1451.391	0.880	0.993	0	154.898
3	1409.298	0.710	0.866	0	145.571
4	1433.001	0.898	0.986	0	177.554
5	1425.372	0.878	0.918	0	195.407
6	1426.434	0.918	0.957	0	186.020
7	1316.693	0.468	0.278	0	71.075
8	1420.011	0.933	0.975	2	193.949
9	1435.581	0.853	0.782	1	225.048

Table 3: Output from the WLS regression with group K-fold cross validation (K=10). Model m7 exhibits the lowest mean RMSE across folds but also has a very low R2, indicating poor data fit. The last two columns display the maximum absolute significant regression coefficient value and its corresponding predictor dimension. For all models except the last two, the largest absolute regression coefficient was found in the first dimension (dimension 0).

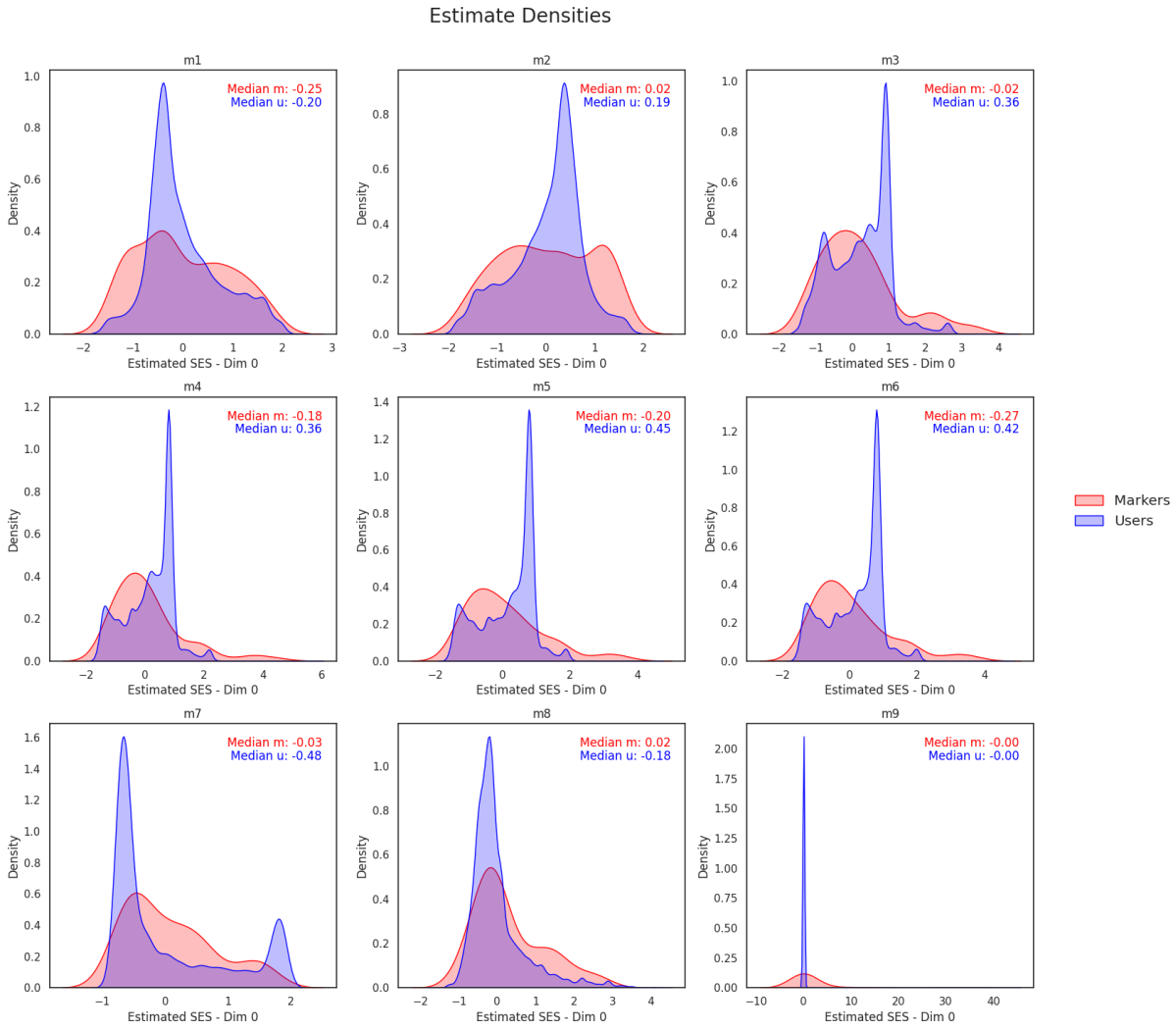


Figure 6: KDE plots illustrating the densities of marker and user positions in the first dimension across all models.

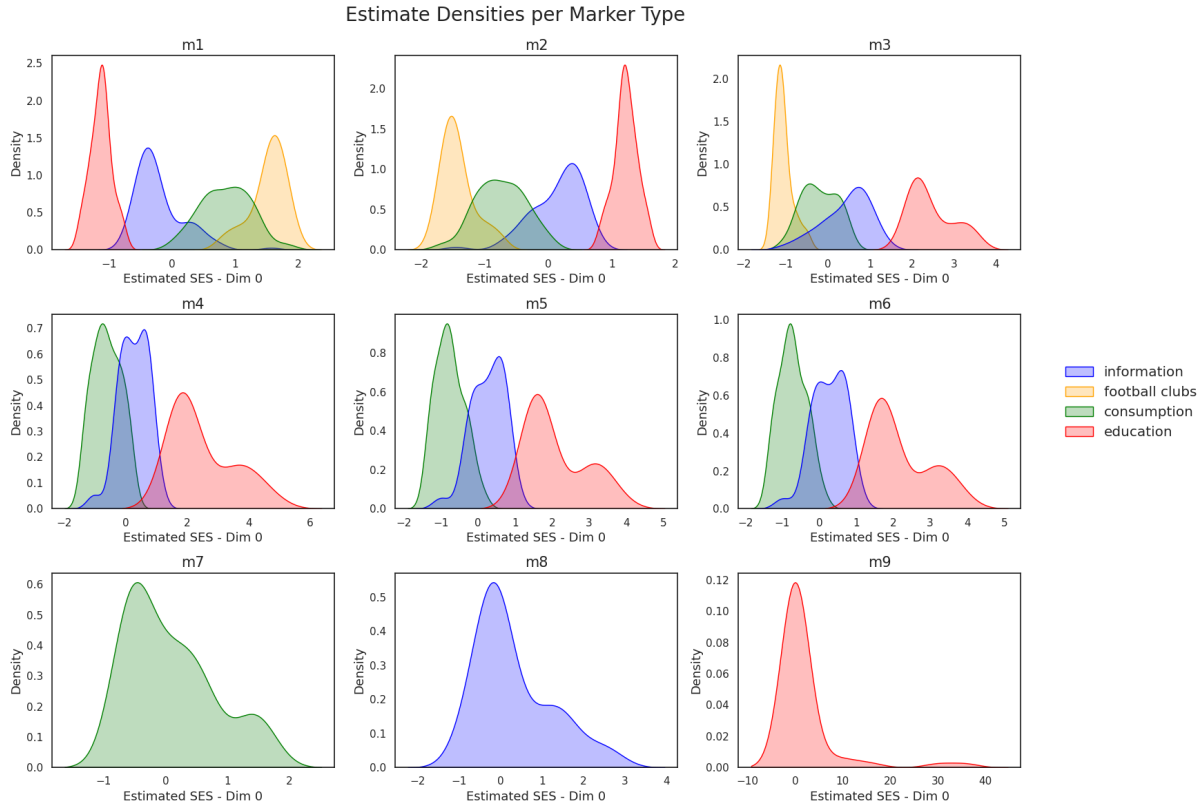
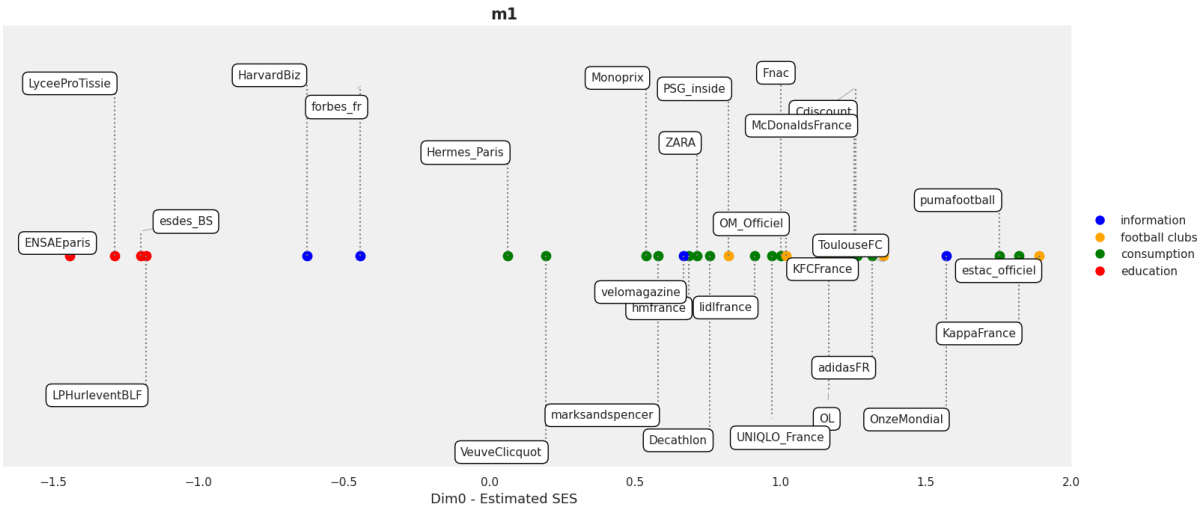


Figure 7: KDE density distributions of marker positions in the first dimension across all nine models. Note the consistent order of marker types across models, indicating the preservation of their relative positions.



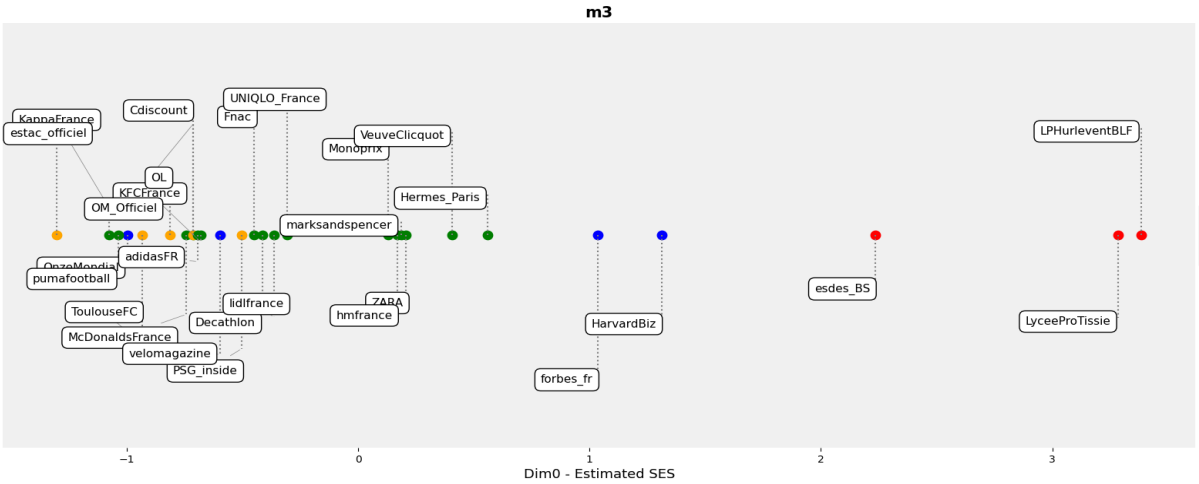


Figure 8: Relative positions of selected markers in the first dimension for m1 and m3. Educational markers are positioned the furthest from the origin in both models, suggesting their distinction. Overall, a similar order for the selected common markers is found for the two.

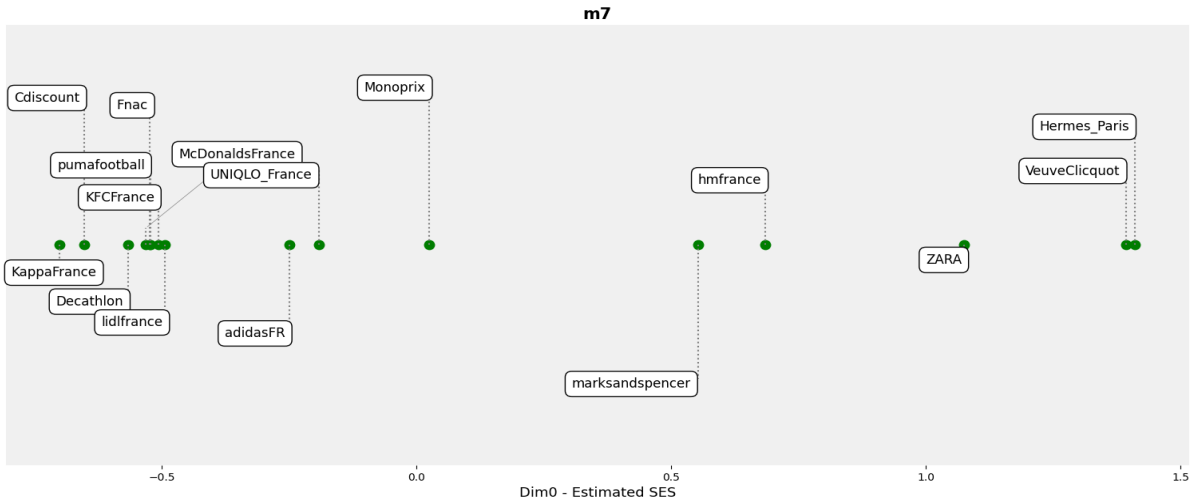


Figure 9: The figure shows the positions of a selection of markers for m7. Relative positions of markers are similar to the corresponding markers in m1 and m3, but a more clear grouping of the consumption markers is observed here. Luxury brands are still positioned the furthest from the origin, suggesting their distinction from the rest of the consumption markers in m7.

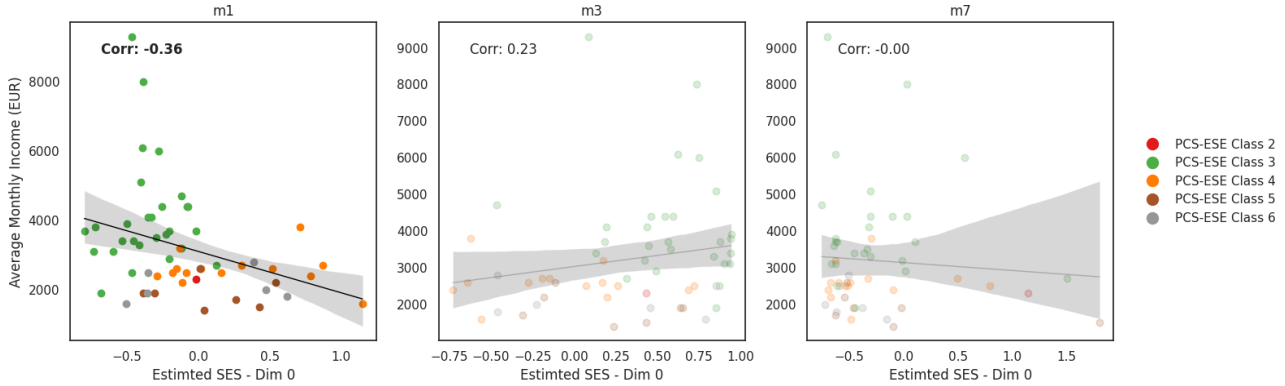


Figure 10: Correlation between median estimates per PCS-ESE job group in the first dimension and income for the selected models. Points are coloured according to the main groups of the PCS-ESE scale, with each point representing the median coordinate of a specific PCS-ESE group in the validation data. Only m1 displayed a significant correlation coefficient.

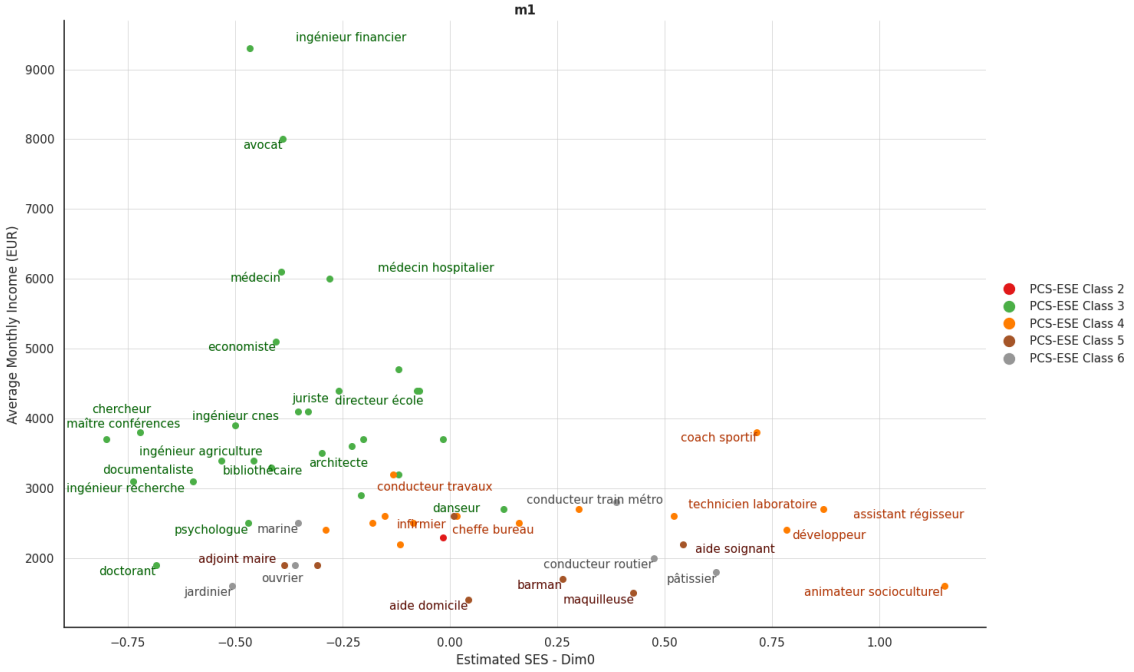


Figure 11: Scatter plot of SES estimates in m1 for selected job titles in the validation data, plotted against their associated average monthly income per job group in EUR. Points are coloured according to the main groups of the PCS-ESE scale, with each point representing the median coordinate of a specific PCS-ESE job group. Labels are added for certain job titles within PCS-ESE groups, though they do not represent all job titles within those groups. For a comprehensive list of all job titles associated with each PCS-ESE group, refer to the Appendix (A.5).

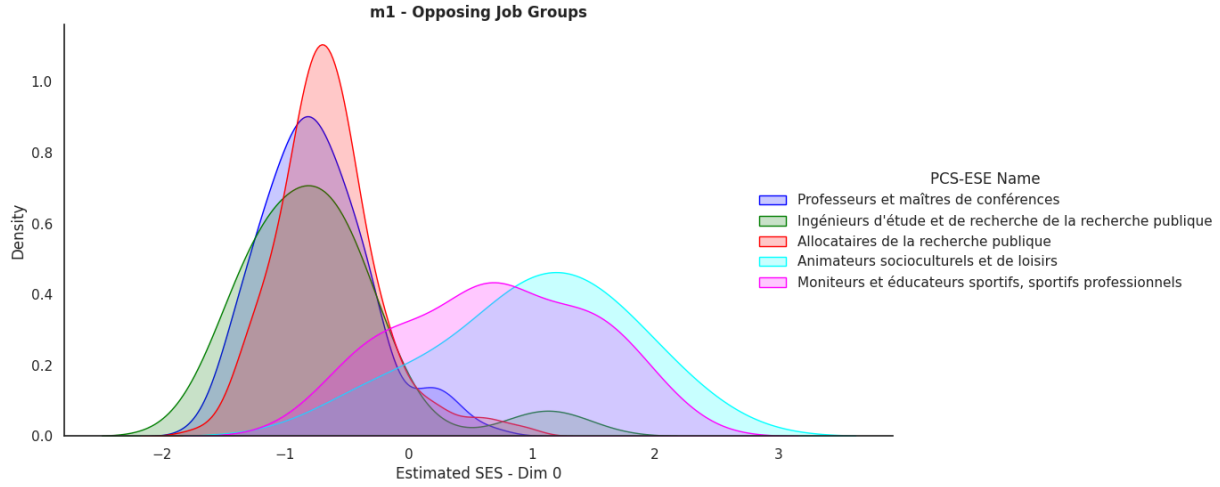


Figure 12: KDE density plot showing the distributions of opposing job groups in m1. Job groups with the highest estimates (negative end of the scale) exhibit narrower distribution intervals compared to lower scoring groups, indicating greater variability within the lower scoring groups.

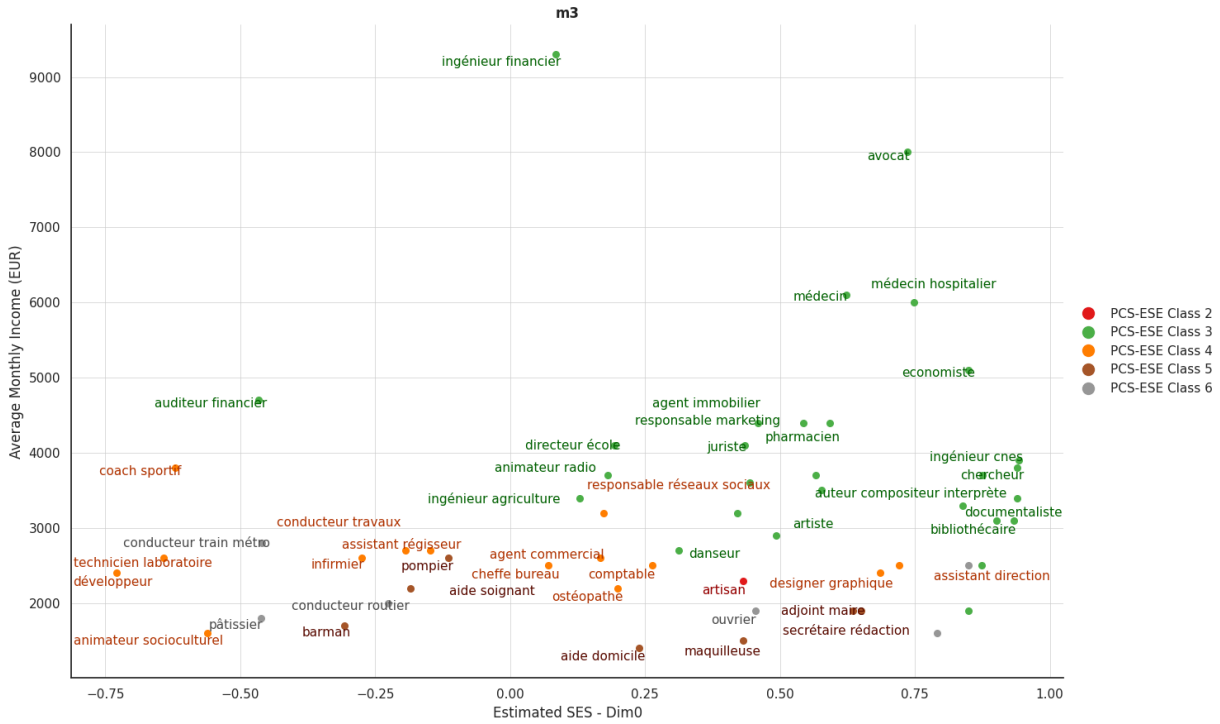


Figure 13: Scatter plot of median SES estimates in m3 for selected job titles in the validation data, plotted against the average income per job group in EUR. The colouring of points and selection of job titles follow the same logic as described for Figure 12.

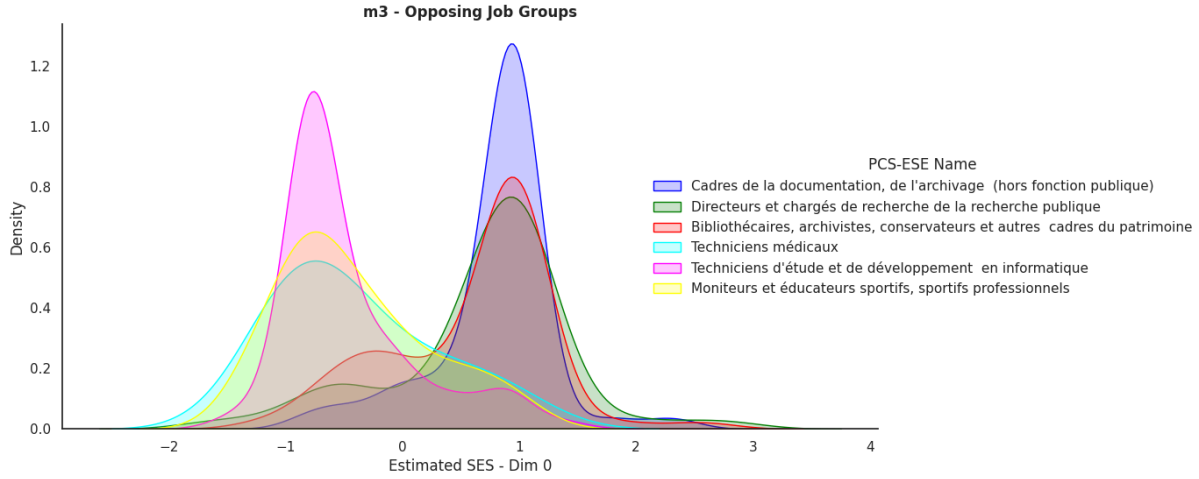


Figure 14: KDE density plot showing the distributions of opposing job groups in m3. Similar to m1, job groups with higher SES estimates exhibit narrower distributions compared to those with lower estimates. However, both groups show larger tails, indicating greater variability.

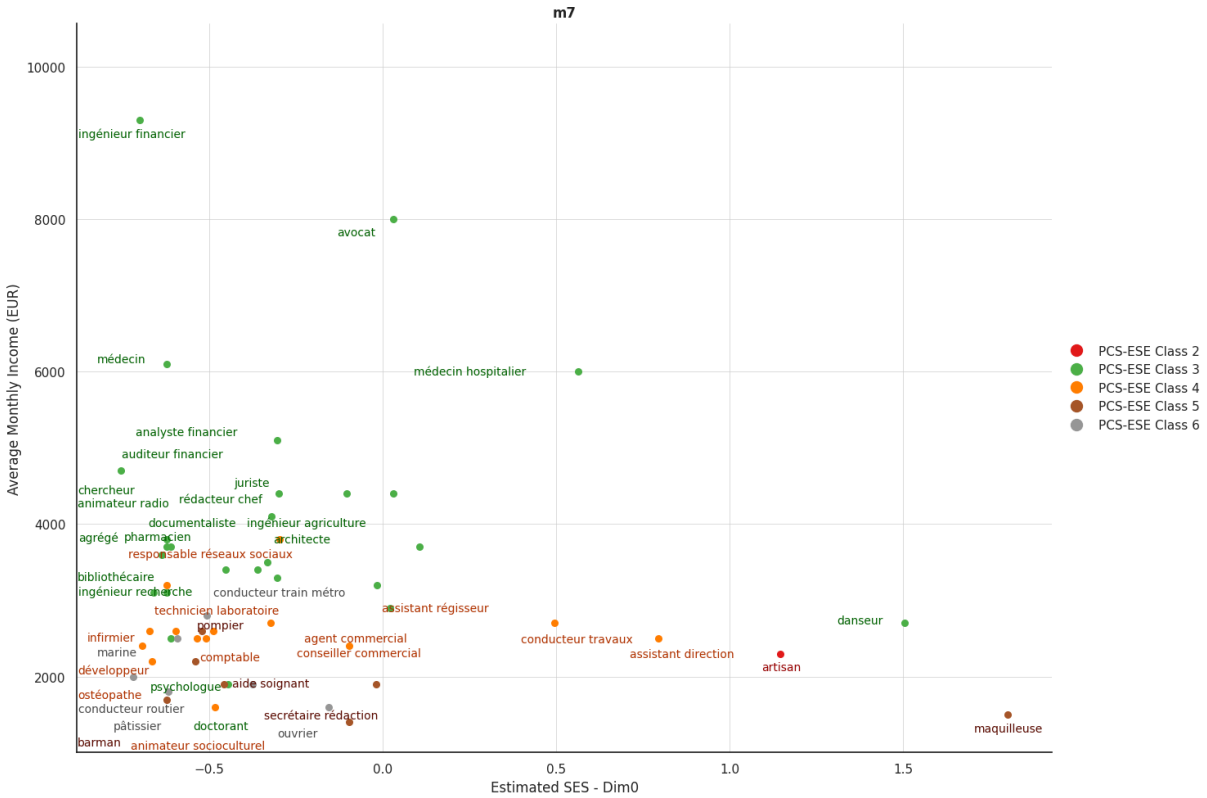


Figure 15: Median SES estimates in m7 for selected job titles plotted against the average income per job group in EUR. The colouring of points and selection of job titles follow the same logic as described for previous similar plots. A less distinct relationship is observed between median SES estimates for the job groups and their associated income compared to m1 and m3. Most job groups are clustered within a similar SES estimate range.

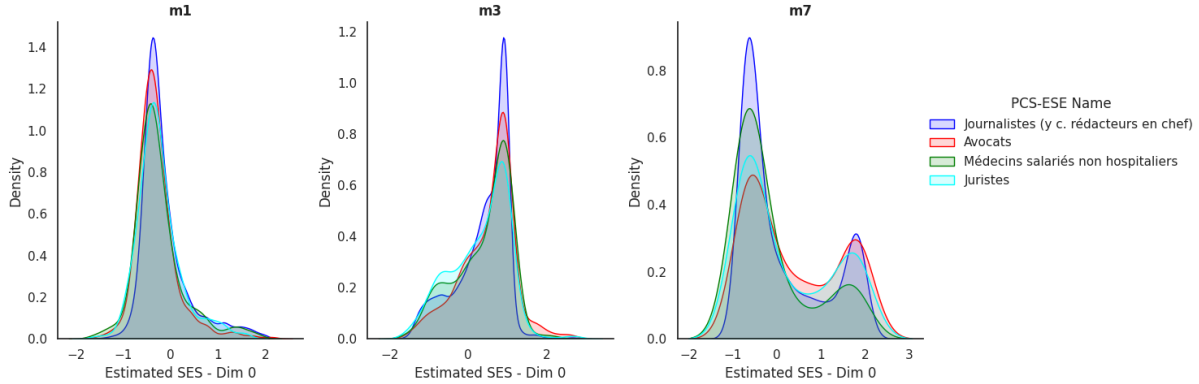


Figure 16: KDE plot of outlier job groups identified for having the widest estimate range and the highest number of individual estimates outside the main estimate range for their respective groups. Each group exhibits distinct main peaks but displays wide tails, suggesting significant deviation of certain users from the central distribution.