Supervised and unsupervised learning with Johnson and Johnson ticketing database

Ignacio Palma, Jairo Melo, Mahboob Jamil and Vikram Khade

22 March 2019

1 Introduction

2 Clustering

Partitioning Around Medoids (**PAM**) technique is used for clustering.

I use features :

levelN, priorityN, impactN, app_category , res_category, region, ndays, prod_line

Issue is - level2 dominates. The dataset is imbalanced especially with respect to L3 which happens to be very important.

Solution : Undersample L1 and L2.

set.seed(123)
indx2 = sample(which(sdata$levelN == 2), round(0.2 * sum(sdata$levelN == 2)),replace=FALSE)
set.seed(123)
indx1 = sample(which(sdata$levelN == 1), round(0.8 * sum(sdata$levelN == 1)),replace=FALSE)

underDF = rbind(sdata[indx1,], sdata[indx2,])
underDF = rbind(underDF,sdata[which(sdata$levelN == 3),])

The underDF is more balanced than the original dataset. Though nrows = 6034 only.

Without undersampling, all clusters tend to be dominate by either L1 or L2. Our main interest is L3.

gower

3 Conclusion

Clus #	level	Priority	Impact	app_category	res_category	Region	Prod_line	ndays
Cluster 17	L3	P1 & P2	I1	Software	Data Issues	1028	Line1	Mixed
Cluster 18	L2 & L3	P2	I1	Application	Configuration	1007	Line2	Mixed
Cluster 15	L2	P2	I1	Software	Job Failure	1028	Line1	Mixed

This identified segment should be looked into by J&J to formulate strategies to push these tickets into L1 category to save costs.

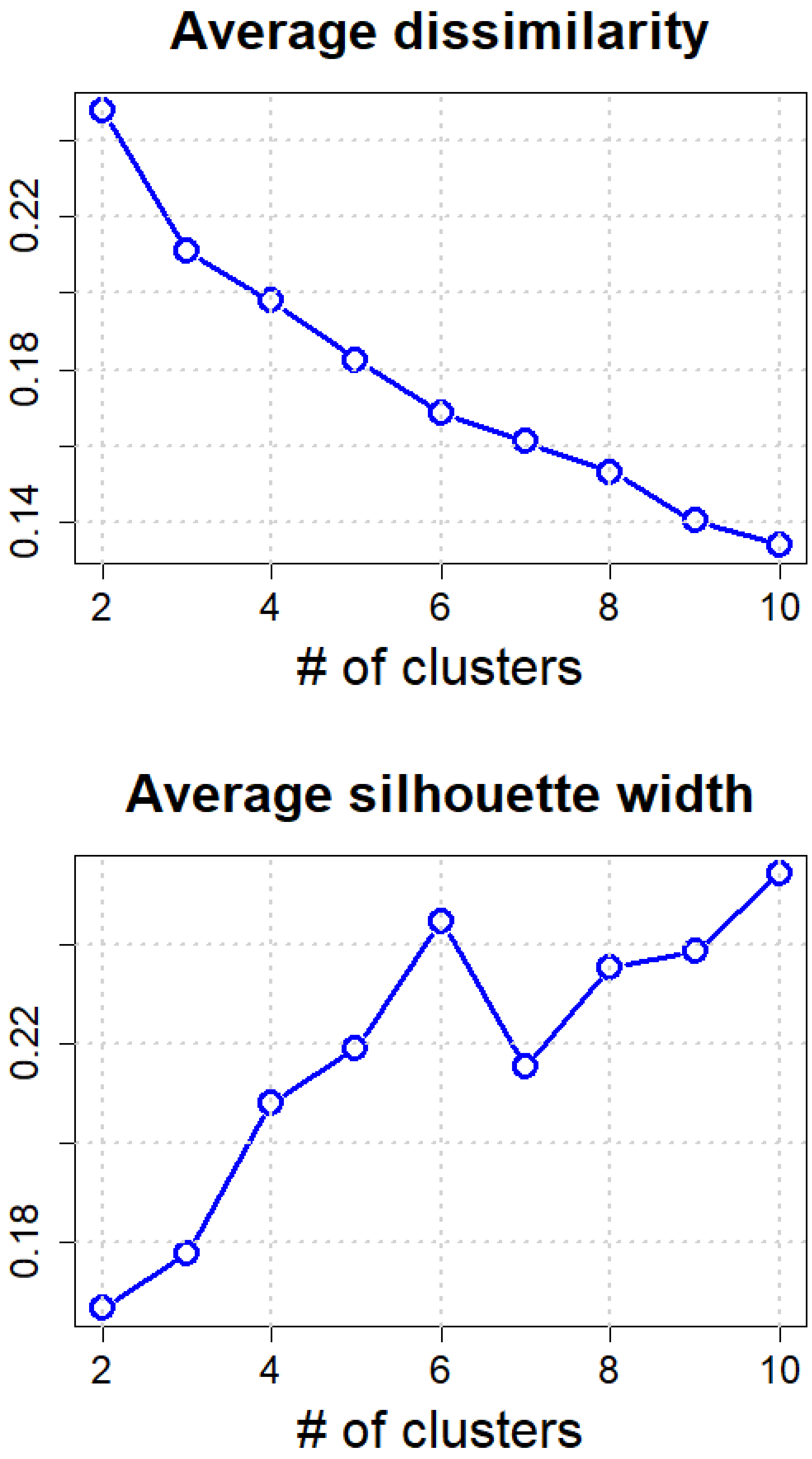


Figure 1: **For full sample of 17423 datapoints.**

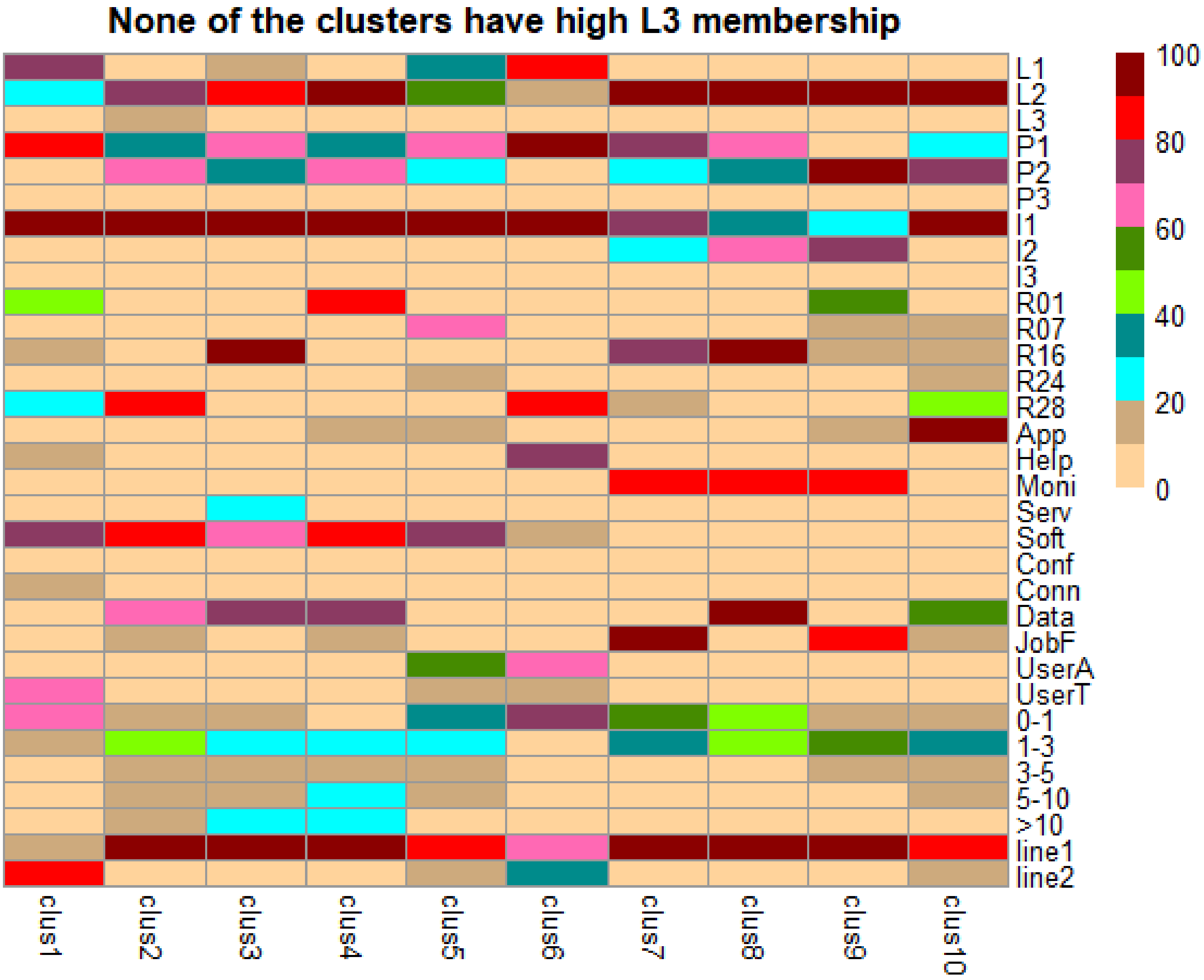
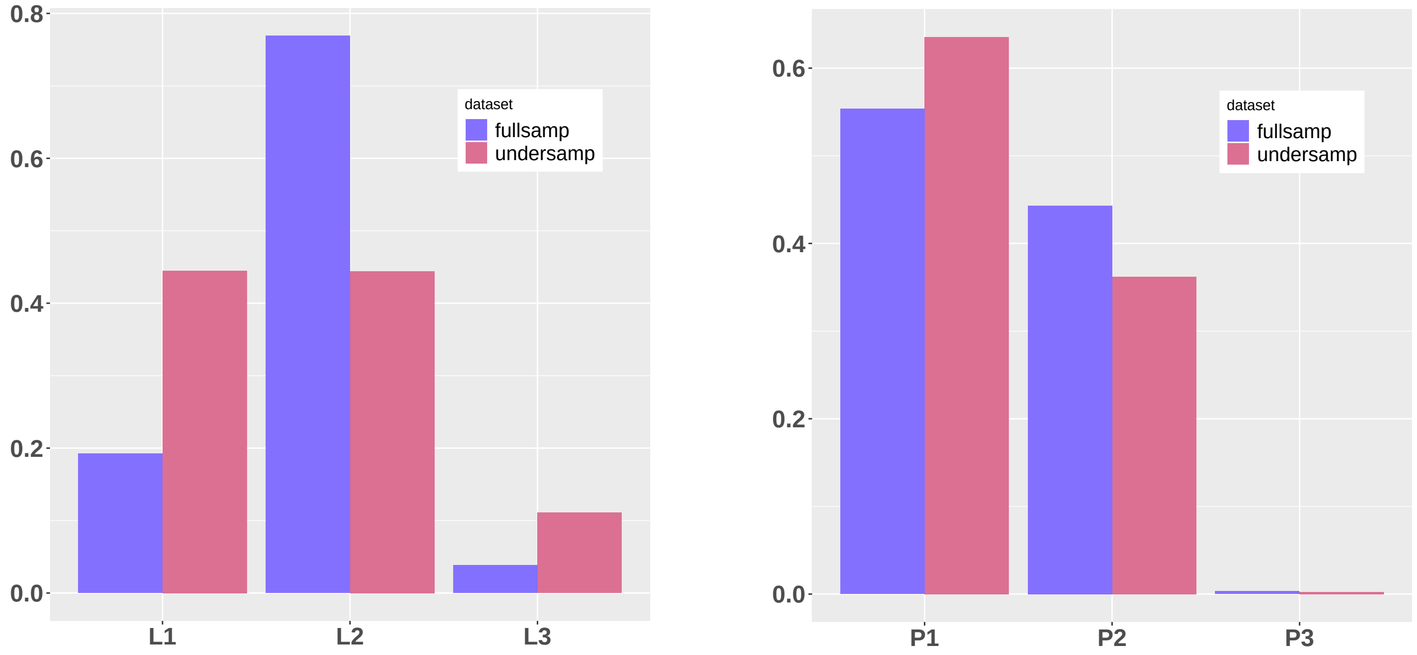


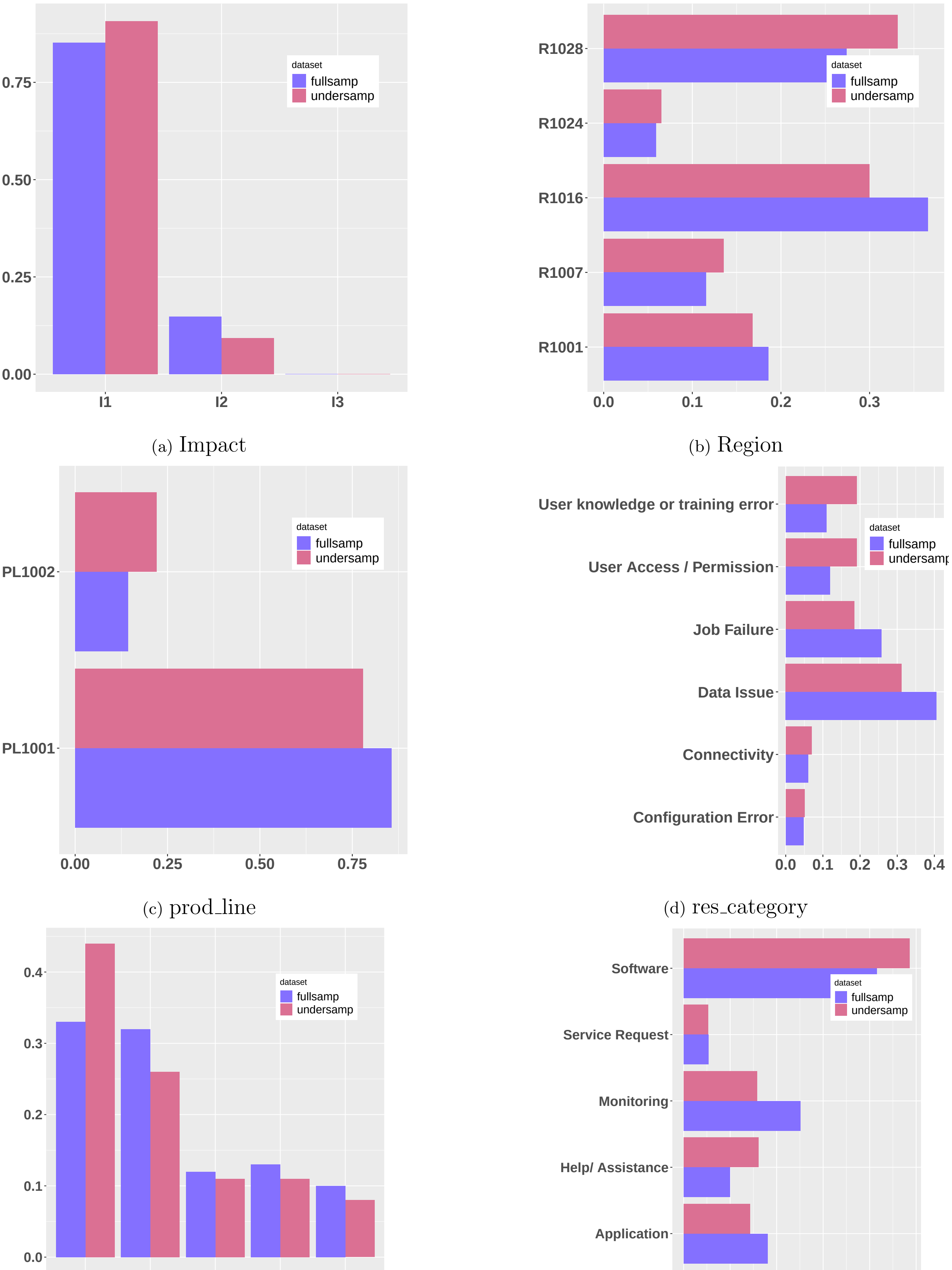
Figure 2: **For full sample of 17423 datapoints : the 32 rows are the total number of levels of all features used for clustering. Each column shows the membership of a particular cluster in percentage. The percentages for a particular feature add up to 100 for each cluster. None of the clusters have a high membership of level L3. This is because L3 is a small minority in the dataset compare to L1 and L2. Most clusters are dominated by L2 which makes up about 78% of the dataset for the level feature.**



(a) L2 dominates in full sample dataset (blue color).

(b) Priority

Figure 3: **The majority levels L2 and L1 are undersampled. The distribution of the undersample is shown in the reddish color. As a side effect distribution in other features is changed. Priority is shown in the right side panel. The overall shape of priority after undersampling remains the same.**



(a) Impact

(b) Region

(c) prod_line

(d) res_category

(e) ndays

(f) app_category

Figure 4: **Undersampling the level feature changes the distribution of other features.**

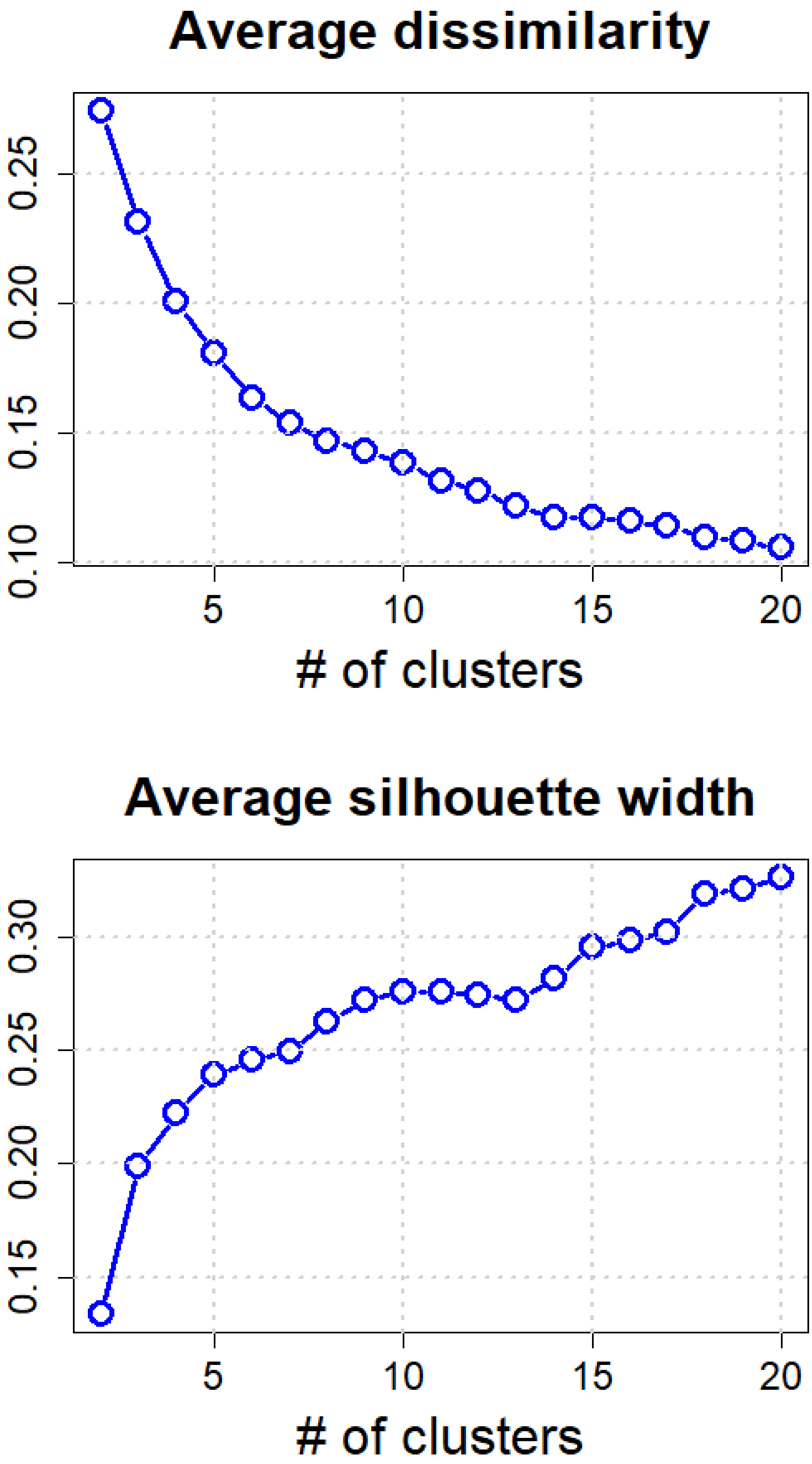


Figure 5: **For undersample of 6023 datapoints : Average silhouette width continues to increase with value of K. However it is hard to analyse more than 20 clusters. Hence k=20 is chosen for further analysis.**

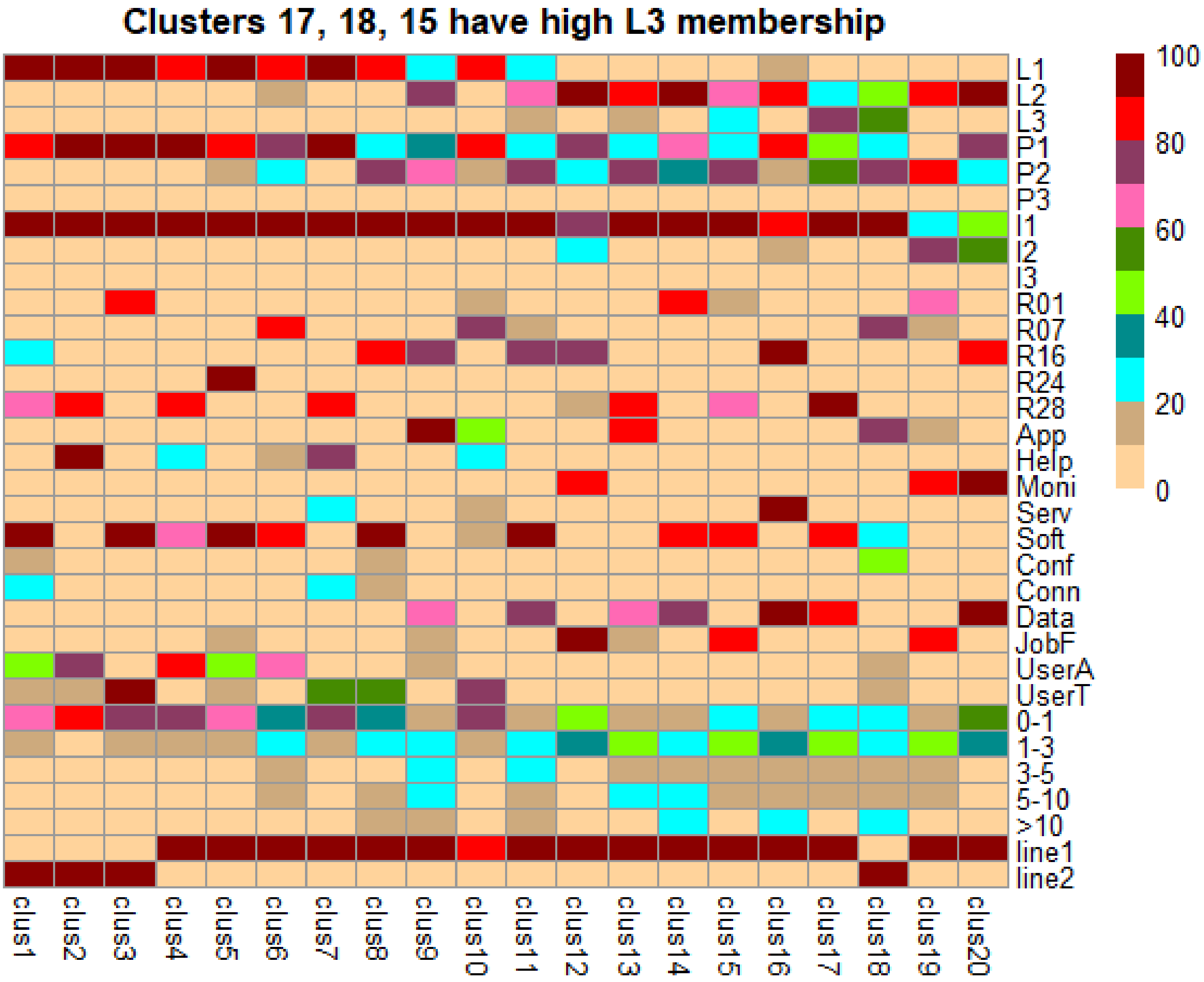
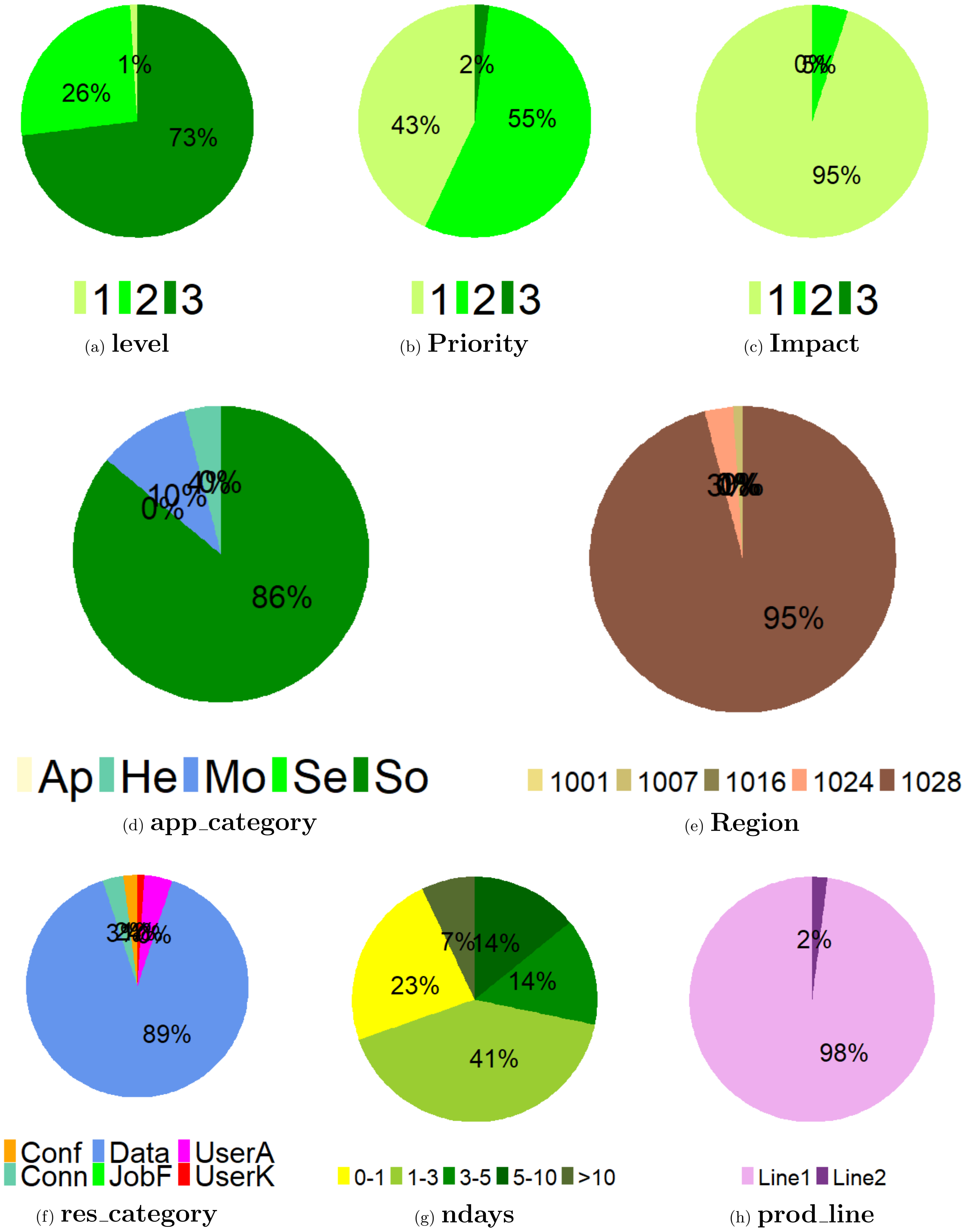


Figure 6: **For undersample of 6023 rows. Undersampling improves the membership of L3 level.**

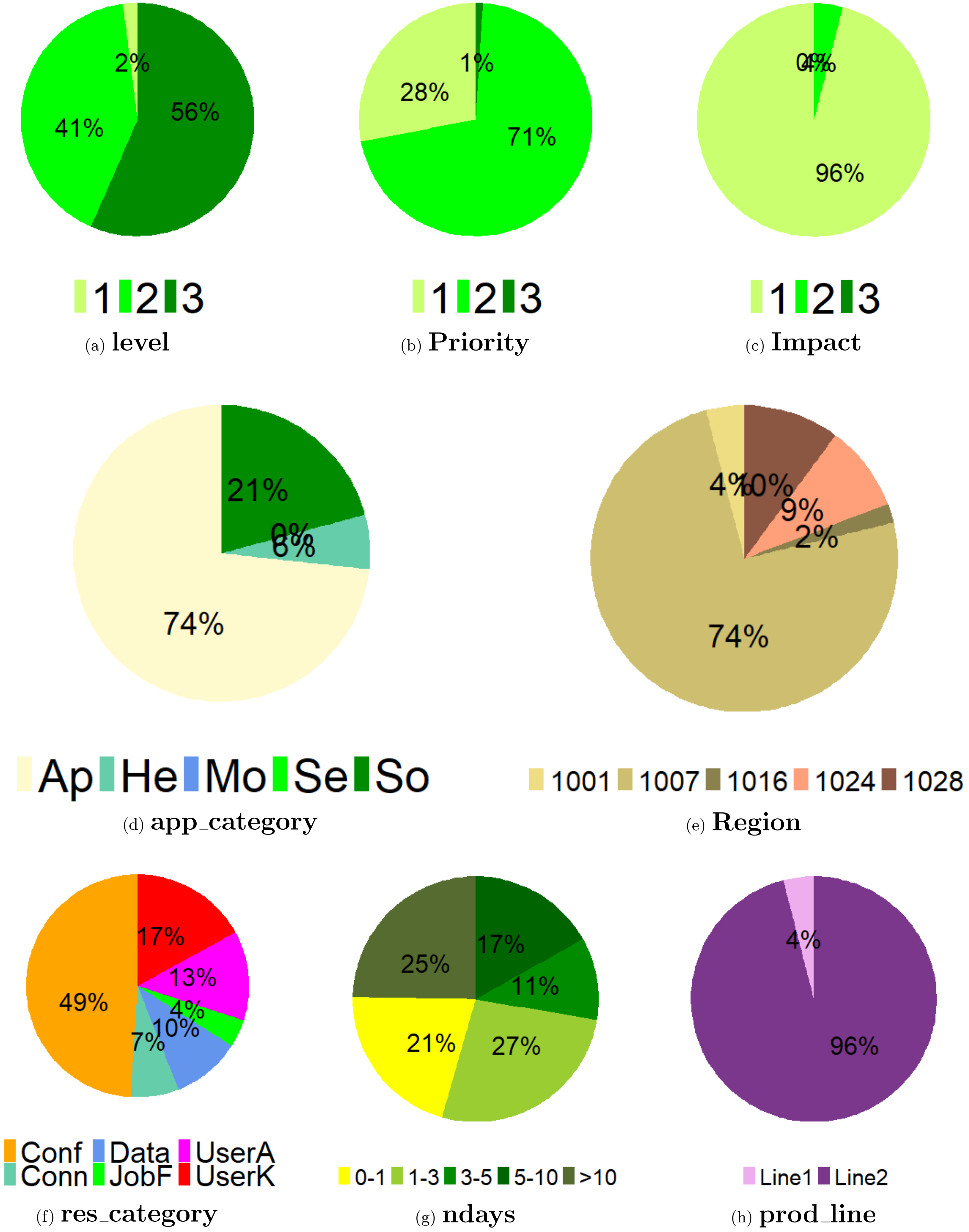


(a) **level** (b) **Priority** (c) **Impact**

(d) **app_category** (e) **Region**

(f) **res_category** (g) **ndays** (h) **prod_line**

Figure 7: **CLUSTER 17**: Level is dominated by L3. These tickets primarily are about *software* and *data issues*. Also an overwhelming 95% of these tickets come from region 1028. Features Impact, Region and prod_line have very *pure* membership in this cluster. ndays is quite mixed.



(a) **level** (b) **Priority** (c) **Impact**

(d) **app_category** (e) **Region**

(f) **res_category** (g) **ndays** (h) **prod_line**

Figure 8: **CLUSTER 18**: Level is dominated by L3 and L2. These tickets primarily are about *application* and *configuration issues*. Also a majority 74% of these tickets come from region 1007. Features Impact and prod_line have very *pure* membership in this cluster. ndays is quite mixed. In contrast to cluster 17, production line 2 dominates the prod_line feature.

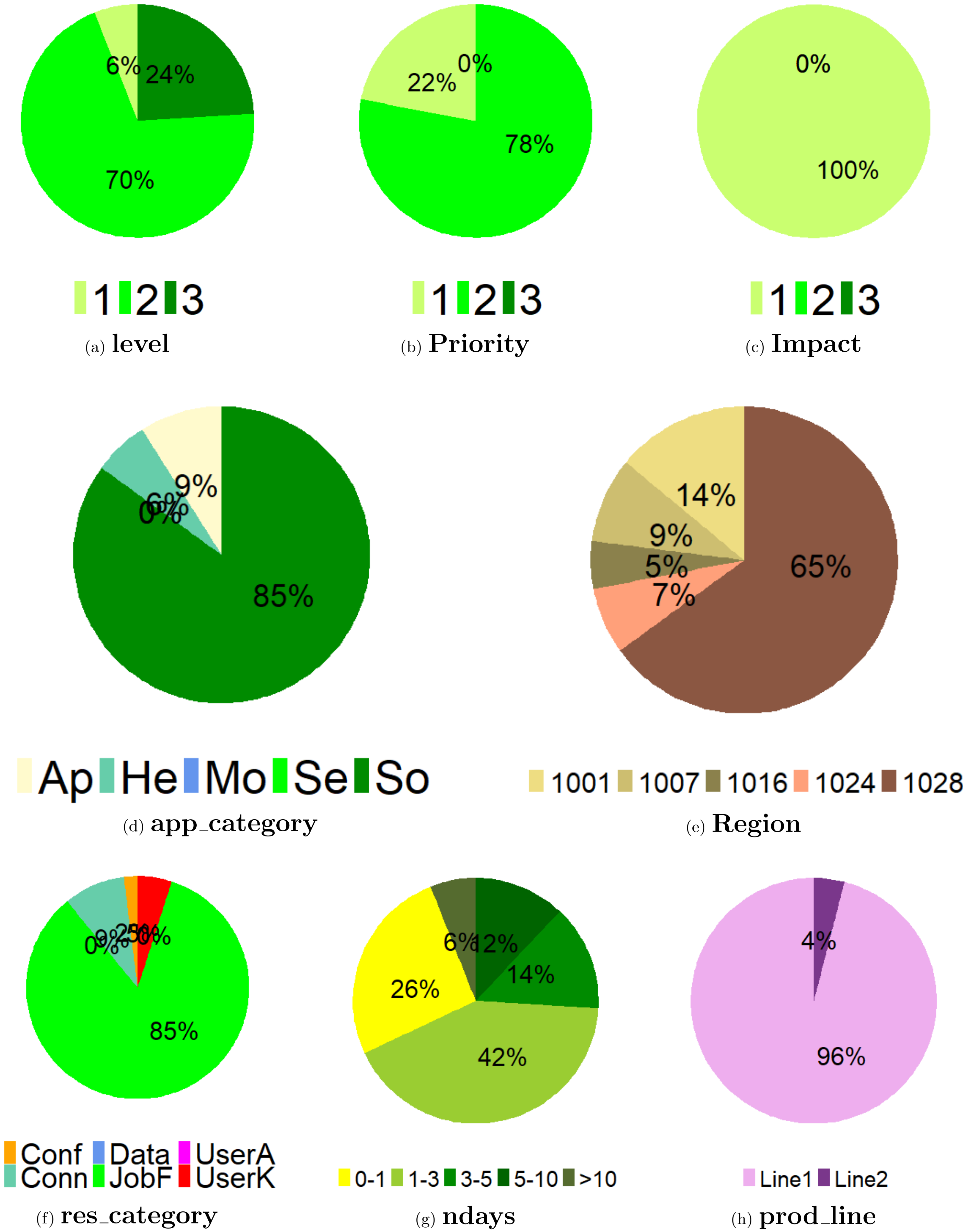


Figure 9: **CLUSTER 15**: Level is dominated by L2 and L3. These tickets primarily are about *software* and *job failure*. Also a majority 65% of these tickets come from region 1028. Features Impact and prod_line have very *pure* membership in this cluster. ndays is quite mixed.