Multiplicity and clustering in Taurus star forming region
II. From ultra-wide pairs to dense NESTs*

Isabelle Joncour1,2, Gaspard Duchène1,3, Estelle Moraux1, and Frédérique Motte1

1 Univ. Grenoble Alpes, CNRS, IPAG, 38000 Grenoble, France
e-mail: isabelle.joncour@univ-grenoble-alpes.fr
2 Department of Astronomy, University of Maryland, 20742 College Park, MD, USA
e-mail: joncour@galia.astro.umd.edu
3 Astronomy Department, University of California, 94720-3411 Berkeley, CA, USA

Received 18 March 2018 / Accepted 31 August 2018

ABSTRACT

Context. Multiplicity and clustering of young pre-main sequence stars appear as critical clues to understand and constrain the star formation process. Taurus is the archetypical example of the most quiescent star forming regions that may still retain primeval signatures of star formation.

Aims. This work identifies local overdense stellar structures as a critical scale between wide pairs and loose groups in Taurus.

Methods. Using the density-based spatial clustering of applications with noise (dbscan) algorithm, and setting its free parameters based on the one-point correlation function and the k-nearest neighbor statistics, we have extracted reliably overdense structures from the sky-projected spatial distribution of stars.

Results. Nearly half of the entire stellar population in Taurus is found to be concentrated in 20 very dense, tiny and prolate regions called NESTs (for Nested Elementary STructures). They are regularly spaced (∼2 pc) and mainly oriented along the principal gas filaments axes. Each NEST contains between four and 23 stars. Inside NESTs, the surface density of stars may be as high as 2500 pc⁻² and the mean value is 340 pc⁻². Nearly half (11) of these NESTs contain about 75% of the class 0 and I objects. The balance between Class I, II, and, III fraction within the NESTs suggests that they may be ordered as an evolutionary temporal scheme, some of them getting infertile with time, while other still giving birth to young stars. We have inferred that only 20% of stars in Taurus do not belong to any kind of stellar groups (either multiple system, ultra wide pairs or NESTs). The mass-size relation for stellar NESTs is very close to the Bonnor–Ebert expectation. The range in mass is about the same as that of dense molecular cores. The distribution in size is bimodal peaking at 12.5 and 50 kAU and the distribution of the number of YSOs in NESTs as a function of size exhibits two regimes.

Conclusions. We propose that the NESTs in their two size regimes represent the spatial imprints of stellar distribution at birth as they may have emerged within few millions years from their natal cloud either from a single core or from a chain of cores. We have identified them as the preferred sites of star formation in Taurus. These NESTs are the regions of highest stellar density and intermediate spatial scale structures between ultra-wide pairs and loose groups.

Key words. methods: statistical – binaries: visual – open clusters and associations: individual: Taurus – ISM: structure – stars: formation

1. Introduction

The Taurus star-forming region is the archetype of a quiescent and sparse star distribution associated to a low surface stellar density (ρ ∼ 0.8 pc⁻², ~5 deg⁻²), spread over an area of ≈420 pc² on the sky for the central part in Taurus. A high-multiplicity fraction and loose spatial clustering are key features in this region. On the clustering side, it is now well established that the stellar population of Taurus is distributed in a few groups of 1.5–2.5 pc size (Jones & Herbig 1979; Gomez et al. 1993; Kirk & Myers 2011). On the multiplicity side, the companion frequency of young stars in Taurus is generally twice that of field stars (Duchène & Kraus 2013). Moreover the presence of most probably coeval ultra-wide binaries appears to be a crucial outcome of the star formation process in that region (Joncour et al. 2017; hereafter Paper I), a conclusion that also applies to the Perseus region for the embedded young stellar objects (YSOs; Sadavoy & Stahler 2017). The higher order multiplicity property of these ultra-wide pairs (UWPs) in Taurus suggests a close link between the spatial scales associated with multiple systems and with UWPs, ≤1 kAU and ≤60 kAU respectively. Indeed, the UWPs population appears to extend the traditional multiple system regime by almost two orders in magnitude in spatial scale. In Paper I, we have proposed that this property illustrates the remaining imprints of a molecular core or clump fragmentation cascade scenario that gave birth to young stars. Many of the UWPs in Taurus are made of two multiple systems, and are therefore “small stellar groups” of up to five members. However, the method used to identify UWPs was based on the mutual nearest neighbor property, which could only identify pairs. Higher order multiplicity (groupings of three or more separate systems) were thus not addressed in Paper I. We now wish to detect local stellar groups (overdensities) of any

---

*A Table A.1 is only available at the CDS via anonymous ftp to cdsarc.u-strasbg.fr (130.79.128.5) or via http://cdsarc.u-strasbg.fr/viz-bin/qcat?J/A+A/620/A27

Open Access article, published by EDP Sciences, under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
order, and thus need to use a different methodology. We have adopted a clustering algorithm, which is appropriate for this task.

In continuity with Paper I, the present study aims at analyzing spatial structures within the Taurus stellar population at intermediate length scales. In order to reach this objective, we propose a new methodology based on bottom-up approach using local physical properties to uncover explicitly spatial structures at this scale with a high level of confidence. The paper is organized as follow: in Sect. 2, we describe the methodology that allows the identification with high reliability of the new spatial structures that we called NESTs (Nested Elementary Structures) described in the Sect. 3. In Sect. 4, we characterize those structures and highlight their fundamental properties, while in Sect. 5, we discuss further properties and the nature of the NESTs and give our conclusions in Sect. 6.

2. Method

2.1. Data

We used the same Taurus catalog as described in Paper I. It contains 338 members of Taurus taken from the full census of members down to 0.02 $M_\odot$ (Luhman et al. 2010) that we complement by the multiplicity information, with a high degree (67%) of completeness. Following the seminal work of Larson (1995), we aim at distinguishing clustering from multiplicity. To this end, we grouped together all stars that are separated by less than 1 kAU to form a single entity and simply called them “multiple systems”. While this threshold is somewhat arbitrary, we selected it in Paper I based on two complementary arguments. First of all, it is the lower threshold that defines by consensus wide binaries with separation larger than 1 kAU (Reipurth & Mikkola 2012; Tokovinin 2017). Second, this threshold is close to the beam (seven arcsec) of the Spitzer-MIPS instrument, which is used in the classification of young stellar objects, particularly the most embedded ones. Therefore, the census of neighbor stars is completely independent of the Class of the Taurus members for all separations larger than this threshold.

We note that Kraus et al. (2017) have presented a more recent catalog that includes a new distributed population of diskless stars. The newly identified members are distributed over a broader footprint and belong primarily to the dispersed stellar population, which is typically older than the “classical” stellar population of Taurus. Therefore leaving this population aside does not significantly affect our conclusions surrounding local overdense stellar regions and we retain the catalog from Paper I for our analysis.

2.2. Main approaches to identify stellar subgroups

The typical tools that are used to study the stellar spatial distribution may be subdivided in two main categories. The first focuses on spatial analyses in a global sense, aiming at characterizing the sources distribution as a whole (e.g., the two-point correlation function is used to evaluate the degree and the regimes of clustering, and the one-point correlation function to probe the binary regime range). The second approach aims at extracting (sub-)structures as topological entities using clustering algorithms, to further characterize them and derive their geometrical and physical properties. This task of finding clusters in a set of points has a long history in the field of applied mathematics and computer science. But intuitively, clusters can be seen simply as regions of enhanced stellar density with respect to their surroundings. Despite this simple definition, decades of dedicated and still on-going work show that there exist several methods to identify clusters (see Appendix C for a short review and references therein). Each of these methods has its pros and cons, and the optimal algorithm depends heavily on the type of data at hand and on the scientific goals.

The global distribution analysis and the substructures identification are complementary studies and, in some cases, the same tool can be used in both approaches. In particular, this is the case of the Minimum Spanning Tree (MST; Kruskal 1956; Prim 1957; Dijkstra 1960; Zahn 1971). This method is a graph procedure that connects all points in an ensemble through a simple path such that the total distance length of its edges is minimal. The method then generates clusters by deleting the longest individual segments of the MST above a heuristically chosen threshold, such that clusters are defined as the remaining connected subgraphs. In the field of astronomy, the MST method was first applied in extragalactic astrophysics by Barrow et al. (1985). Over the last decade, it is widely used in the context of young star clusters to identify subgroups of young stars, although alternate probabilistic techniques have been proposed recently (e.g., Kuhn et al. 2014).

Gutermuth et al. (2009) have proposed an empirical procedure to identify a critical MST length threshold above which the MST edges may be removed to identify clusters (plus usually a user-based condition on the minimum objects that must be located in the subgroups to be considered as such). This technique was notably used by Kirk & Myers (2011) to study the stellar groups properties in four star-forming regions (Taurus, Lupus III, Cha I, and IC 348).

The MST method can also be used to define useful parameters to quantify the degree of sub-structuring, such as the $Q$ ratio of the average branches length of the MST over the mean separation of all pairs of stars (Cartwright & Whitworth 2004; Schmeja et al. 2008; Wright et al. 2014; Parker 2014), or to investigate mass segregation (Allison et al. 2009). A comparison and discussion on different methods assessing a global measure of the sub-structuring degree and the mass segregation can be found in Schmeja (2011), Küpper et al. (2011), and Parker & Goodwin (2015). The strength of the MST method is that it is a non-parametric technique, but its major drawbacks are inherent to all single linkage techniques, namely the artificial chaining effect, and the impossibility of handling noise and outliers.

2.3. dbscan

In this work, we have chosen to identify spatial features at a given scale using local density properties and a connectivity rule to link together adjacent stars having similar stellar density neighborhood. The spatial structures thus identified are defined as connected ”density cluster”, following previous graph-principled work (Hartigan 1975). To efficiently identify density clusters, we have chosen the non-parametric, one-level dbscan (Density-Based Spatial Clustering Applications with Noise) clustering algorithm developed in the Knowledge Discovery Database field (Ester et al. 1996, see Appendix B). dbscan has several advantages over other clustering algorithms (see Appendix C): the partition of stars within density components is unique, it allows the detection of clusters of arbitrary shape and size at a global scale from local requirements, and it is the only clustering algorithm that explicitly labels noise and outliers. While other clustering algorithms
induce a complete partitioning of the ensemble, dbscan proposes partial clustering. Thus, unlike single linkage and MST algorithms, it is resistant to noise and outliers. The key algorithmic idea of dbscan is to incrementally group stars, provided that (1) they are direct or indirect neighbors, and (2) the stellar density neighborhood of neighbors satisfied the selected criteria. dbscan directly searches for connected dense regions in space separated by local stellar density drops.

In dbscan, the “density cluster” $C_{\epsilon,N_{\text{min}}}$ sets are determined by the choice of two local free parameters: a distance $\epsilon$ and a number of stars $N_{\text{min}}$. Stars are grouped together if they satisfy two conditions: (1) all stars within a set $C_{\epsilon,N_{\text{min}}}$ have a minimum number of stars $N_{\text{min}}$ within a radius $\epsilon$, (2) each star within a set $C_{\epsilon,N_{\text{min}}}$ is connected to any other star of the same set by a sequence of neighbors separated at most by a distance $\epsilon$. With these properties, a density cluster $C_{\epsilon,N_{\text{min}}}$ is said to be maximal among connected sets, i.e., $C_{\epsilon,N_{\text{min}}}$ is not contained in any larger density cluster defined at the same local $\epsilon$ level. As we explain next, the values of $\epsilon$ and $N_{\text{min}}$ that we use in this study have been determined based on the confidence level that local structures are distinct from random fluctuations.

### 2.4. Selecting the algorithm parameters

The clustering property in Taurus is traditionally characterized using the two-point correlation function (TPCF, Gomez et al. 1993; Larson 1995; Simon 1997; Gladwin et al. 1999; Hartmann 2002; Kraus & Hillenbrand 2008). This method reveals an “elbow” in the 4–40 kAU range, which is interpreted either as the signature of the Jeans instability in cool dense molecular cores (Larson 1995) or as an indication for a quasi-constant surface density of pairs produced by random motions that smooth out primordial stellar lumps (Kraus & Hillenbrand 2008). Either way, the TPCF does not yield a unique plausible value for $\epsilon$ and we thus have sought another approach.

In Paper I, we have introduced the one point correlation function defined as the probability of having a closest star located at $r$ from any chosen star at random. It provides a local analysis over all the range of $r$. This function differs and complements the two-point correlation function, the latter being the probability of having pairs separated by $r$. At small $r$, these two functions describe a same spatial property. At larger $r$, the two-point correlation function gives an “integrated” view of the spatial characterization of the spatial distribution. We defined

$$\Psi(r) = \frac{w_T(r)}{w_G(r)},$$

as an estimator of the one-point correlation function, where $w_T$ is the distribution of nearest neighbor distances in Taurus, and $w_G$ is the same distribution for a random distribution with the same mean surface density.

This $\Psi$ function, which encapsulates the coarsest trends in the stellar spatial distribution of star, reveals three different spatial regimes in Taurus (clustering, inhibition and dispersion). The clustering regime extends over all distances associated with an excess of stars that have a nearest neighbor less than $r_c = 0.1^\circ \approx 0.24$ pc ($\approx 50$ kAU) over a random distribution, as seen in Fig. 1 (adapted from Fig. 4 in Paper I). This introduces a natural benchmark for the local scale around which the value of $\epsilon$ has to be set.

We have endeavored to set the two free parameters $\epsilon$ and $N_{\text{min}}$ of the dbscan algorithm by requiring that local overdense features are detected at a high level of significance ($\alpha = 99.85\%$, i.e., three $\sigma$, if the distribution was to be normal) above random expectation. In Paper I, we have derived the theoretical cumulative distribution $W_k(r)$ of the $k$-nearest neighbor distribution in the case of a 2D random distribution:

$$W_k(r) = 2(\pi\rho)^k/G(k) \cdot r^{2k-1} \cdot \exp(-\rho r^2),$$

where $G(k)$ is the Gamma function, $\rho = 5$ stars deg$^{-2}$ is the mean stellar surface density of Taurus, and $r$ is the distance to the $k$-nearest neighbor. The value of $W_k(r)$ as described in Eq. (2), gives the probability for a random star to have a $k$-nearest neighbor located at a distance of $r$ or less. In turn, $\alpha = 1 - W_k(r)$ represents the degree of significance of an over-density defined by $k+1$ stars. Setting $r_c = 0.1^\circ$ as the relevant local length scale, we see in Fig. 2 that the probability of having a first companion (i.e., two stars within $r_c$) is fairly high, i.e., $W_1(r_c) \approx 0.15$. Similarly, the probability of having 2 companions within $r_c$ is $\approx 0.11$. We must therefore consider the third nearest neighbor cumulative distribution. With $N_{\text{min}} = 3 + 1 = 4$, we must set $\epsilon = 0.11^\circ$. Very close to the previously identified length scale benchmark $r_c$, to achieve the required confidence level. In summary, the two free parameters of the dbscan algorithm have been set based on the nearest neighbor statistics analysis, with a requirement to identify local spatial features with a 99.85% level of significance above random expectancies.

### 3. NESTs detection

#### 3.1. A new type of structures in Taurus

With the parameters $\epsilon$ and $N_{\text{min}}$ set to the values defined in the previous section, we ran dbscan on the Taurus catalog. We have identified 20 distinct stellar overdense structures, which we have termed Nested Elementary STructures (NESTs). Each
NEST shelters between four stars – the minimum number of stars imposed by the method – and 23 stars, with a mean (resp. median) number of eight (resp. six) stars (see Table 1, Fig. 3 and Appendix A). Considered as an ensemble, the NESTs contain nearly half of the entire stellar population in Taurus. Eighteen of these NESTs are located in the three principal gas filaments of the Taurus molecular cloud.

To estimate the size of each NEST, we have first defined its convex hull as the smallest convex set of points that contains all of its members. The polygonal window drawn from this set of points provides an estimate of the minimal area A enclosing all systems within the NEST. We have then evaluated the typical distance of each NEST as $R_H = \langle A/N \rangle^{1/2}$; they range from $\approx 5$ to $\approx 80$ kAU. The average of the first nearest neighbor separation (1-NNS) within all the NESTs, $r_1$, is about 20 kAU (i.e., 0.1 pc). Thus NESTs appear as a new, intermediate, type of structure, on similar physical scales to UWP s but containing more elements, yet significantly smaller and denser than the already identified loose groups.

### 3.2. Reliability of the NESTs

Algorithms designed to identify overdense substructures within an ensemble of objects can sometimes produce spurious groupings and incorrectly split large groups into smaller sub-units. To test the reliability of these substructures, we have first verified their separability (i.e., how distinct NESTs really are from one another) and their degree of dilution (i.e., how significantly does the overdensity stands out relative to the stellar population outside the NESTs).

To test the separability between NESTs, we have compared the internal spacing of stars belonging to a single NEST to their distance to any other NESTs. The evaluation of this separability for each NEST is performed by introducing a separability ratio $\alpha_N = D_N/r_N$, defined as the ratio of the nearest distance ($D_N$) from a star in a given NEST to a star belonging to any other NEST over the mean 1-NNS within the NEST ($r_N$). Large values of the ratio indicate well-separated NESTs. All but 2 NESTs have $\alpha_N \gtrsim 3$, with a median value of $\approx 11$. In other words, stars in a NEST are on average one order of magnitude closer to other members of the same NEST than to stars in other NESTs. NESTs number 10 and 12 are the least separated NESTs, raising the possibility that they are two substructures within a larger one.

In a second test, we have evaluated the degree of dilution of the NESTs relative to the more dispersed population, that is the stellar population outside the NESTs. Specifically, for a given NEST, we have first computed the nearest distance $d_*$ between a NEST member and a star of the dispersed population and we have defined the dilution parameter $\beta_N = d_*/r_N$. Smaller values of $\beta_N$ indicate that it is harder to distinguish a NEST from the surrounding population, as the overdensity of the NEST becomes increasingly marginal. Values for $\beta_N$ range from $\approx 1$ to $\approx 13$, with a median of 4. Therefore, NESTs are not in complete isolation but rather immersed in the more dispersed population.

As Fig. 4 shows, the $\alpha_N$ and $\beta_N$ quantities are positively correlated ($p$-value of $10^{-3}$ based on nonparametric Spearman rank correlation test). Although at the size that is proportional to the mean 1-NNS, we note that the quantities $r_N$ and $\alpha_N$ on the one hand, and $r_N$ and $\beta_N$ on the other, appear significantly anti-correlated ($p$-value of $10^{-3}$ for both quantities). This indicates that the more compact a NEST is, the further it is from the nearest NEST and the more its members are separated from the dispersed population. Taking these correlations together, we conclude that there is a connection between the internal local density of the NEST and the local density of the immediate stellar neighborhood. Consequently, the NESTs and the dispersed population are somehow connected and must be interpreted in a comprehensive model.

### 3.3. Robustness of the NESTs

We now focus on the NESTs robustness. Since their detection is predicated on setting the two local parameters $\epsilon$ and $N_{\text{min}}$, we must explore whether changing the values of these parameters affect the results of our analysis. The results of these tests are illustrated in Fig. 5. In a first test, we have varied $N_{\text{min}}$ while keeping $\epsilon$ at its fiducial value. Increasing this parameter by one ($N_{\text{min}} \rightarrow 5$) automatically eliminates all NESTs that contain only four stellar systems. However, it does not affect the detection of all others (see top panel of Fig. 5). Conversely, all fiducial NESTs were retrieved if $N_{\text{min}}$ is decreased to 3, with the addition of three new features, one located just south of NEST number 17 and the two others in the immediate vicinity of two big NESTs (numbers 2 and 14). Because they contain only three systems, these newer NESTs have a slightly higher probability of occurring from random fluctuations, and we place a 97.9% confidence level on their physical nature.

We have then evaluated the effect of the local radius $\epsilon$ by varying its value by $\pm 10\%$ and $\pm 50\%$ (see central and bottom panels of Fig. 5) while keeping $N_{\text{min}} = 4$. This range of variations is associated with the full recovery of the multiscale stellar structure in Taurus (Joncour et al., in prep.). A decrease of ten percent of $\epsilon$ have had no effect on the NEST identification, while an equivalent increase lead to the identification of a new feature (containing four stars) along filament number 1, south of NESTs numbers 17 and 18, at the same location where a new feature appeared when decreasing $N_{\text{min}}$. Furthermore, the same ten percent increase on $\epsilon$ has driven the merging of the two NESTs numbers 10 and 12 into a single larger NEST. This is a consequence of both NESTs having the lowest values of $\alpha_N$ (see Table 1) indicating that they are...
Table 1. Properties of the NESTs.

<table>
<thead>
<tr>
<th>N</th>
<th>n∗</th>
<th>m∗ (M⊙)</th>
<th>αB (◦)</th>
<th>δB (◦)</th>
<th>ρ̅ (pc⁻¹)</th>
<th>RH (kAU)</th>
<th>a (kAU)</th>
<th>a/b</th>
<th>e</th>
<th>PA (◦)</th>
<th>rN (kAU)</th>
<th>αN</th>
<th>βN</th>
<th>Fil.</th>
<th>GKM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15</td>
<td>8.24</td>
<td>63.62</td>
<td>28.14</td>
<td>70</td>
<td>53.8</td>
<td>112.7</td>
<td>2.60</td>
<td>0.92</td>
<td>119</td>
<td>18.2</td>
<td>19.2</td>
<td>3.5</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>23</td>
<td>13.29</td>
<td>64.67</td>
<td>28.40</td>
<td>48</td>
<td>81.0</td>
<td>155.4</td>
<td>1.82</td>
<td>0.84</td>
<td>117</td>
<td>17.5</td>
<td>20.0</td>
<td>4.4</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>1.84</td>
<td>64.95</td>
<td>27.22</td>
<td>533</td>
<td>10.1</td>
<td>37.4</td>
<td>5.67</td>
<td>0.98</td>
<td>143</td>
<td>18.4</td>
<td>7.9</td>
<td>5.1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td>4.69</td>
<td>65.43</td>
<td>26.98</td>
<td>83</td>
<td>40.5</td>
<td>91.3</td>
<td>2.92</td>
<td>0.94</td>
<td>115</td>
<td>15.9</td>
<td>9.1</td>
<td>4.6</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>4.83</td>
<td>66.74</td>
<td>26.11</td>
<td>2.453</td>
<td>5.3</td>
<td>15.3</td>
<td>3.52</td>
<td>0.96</td>
<td>111</td>
<td>5.0</td>
<td>69.5</td>
<td>13.3</td>
<td>1</td>
<td>–</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>1.65</td>
<td>67.41</td>
<td>26.54</td>
<td>437</td>
<td>11.1</td>
<td>34.6</td>
<td>4.14</td>
<td>0.97</td>
<td>111</td>
<td>15.5</td>
<td>22.2</td>
<td>7.8</td>
<td>1</td>
<td>–</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>2.73</td>
<td>67.40</td>
<td>24.58</td>
<td>75</td>
<td>29.9</td>
<td>49.1</td>
<td>1.40</td>
<td>0.70</td>
<td>133</td>
<td>33.3</td>
<td>6.8</td>
<td>2.4</td>
<td>2</td>
<td>–</td>
</tr>
<tr>
<td>8</td>
<td>6</td>
<td>7.73</td>
<td>67.91</td>
<td>18.20</td>
<td>360</td>
<td>15.0</td>
<td>31.7</td>
<td>2.46</td>
<td>0.91</td>
<td>179</td>
<td>10.8</td>
<td>211.6</td>
<td>6.2</td>
<td>–</td>
<td>4</td>
</tr>
<tr>
<td>9</td>
<td>8</td>
<td>4.11</td>
<td>68.11</td>
<td>24.38</td>
<td>65</td>
<td>40.9</td>
<td>87.0</td>
<td>2.64</td>
<td>0.93</td>
<td>108</td>
<td>21.2</td>
<td>10.6</td>
<td>2.9</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>10</td>
<td>7</td>
<td>3.67</td>
<td>68.25</td>
<td>22.84</td>
<td>45</td>
<td>46.1</td>
<td>89.0</td>
<td>2.35</td>
<td>0.91</td>
<td>115</td>
<td>36.4</td>
<td>1.8</td>
<td>4.2</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>11</td>
<td>6</td>
<td>3.85</td>
<td>68.40</td>
<td>26.23</td>
<td>105</td>
<td>27.8</td>
<td>59.6</td>
<td>2.40</td>
<td>0.91</td>
<td>110</td>
<td>18.8</td>
<td>21.3</td>
<td>9.4</td>
<td>1</td>
<td>–</td>
</tr>
<tr>
<td>12</td>
<td>4</td>
<td>1.61</td>
<td>68.51</td>
<td>22.87</td>
<td>137</td>
<td>19.9</td>
<td>39.8</td>
<td>1.69</td>
<td>0.81</td>
<td>137</td>
<td>28.6</td>
<td>2.3</td>
<td>3.7</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>13</td>
<td>4</td>
<td>2.07</td>
<td>68.89</td>
<td>24.18</td>
<td>324</td>
<td>12.9</td>
<td>34.4</td>
<td>2.93</td>
<td>0.94</td>
<td>159</td>
<td>17.2</td>
<td>16.5</td>
<td>7.7</td>
<td>2</td>
<td>–</td>
</tr>
<tr>
<td>14</td>
<td>13</td>
<td>9.81</td>
<td>68.98</td>
<td>22.90</td>
<td>82</td>
<td>46.4</td>
<td>92.7</td>
<td>2.37</td>
<td>0.91</td>
<td>109</td>
<td>18.3</td>
<td>7.2</td>
<td>3.2</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>15</td>
<td>4</td>
<td>1.46</td>
<td>69.60</td>
<td>26.18</td>
<td>447</td>
<td>11.0</td>
<td>25.5</td>
<td>2.21</td>
<td>0.89</td>
<td>97</td>
<td>17.1</td>
<td>8.8</td>
<td>5.5</td>
<td>1</td>
<td>–</td>
</tr>
<tr>
<td>16</td>
<td>4</td>
<td>0.66</td>
<td>69.76</td>
<td>23.60</td>
<td>795</td>
<td>8.3</td>
<td>17.5</td>
<td>1.85</td>
<td>0.84</td>
<td>165</td>
<td>12.0</td>
<td>34.7</td>
<td>7.3</td>
<td>2</td>
<td>–</td>
</tr>
<tr>
<td>17</td>
<td>5</td>
<td>2.07</td>
<td>69.98</td>
<td>26.03</td>
<td>131</td>
<td>22.8</td>
<td>44.0</td>
<td>1.55</td>
<td>0.76</td>
<td>170</td>
<td>22.3</td>
<td>4.5</td>
<td>4.1</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>18</td>
<td>6</td>
<td>2.56</td>
<td>70.29</td>
<td>25.90</td>
<td>48</td>
<td>41.1</td>
<td>61.6</td>
<td>1.36</td>
<td>0.68</td>
<td>140</td>
<td>28.8</td>
<td>3.5</td>
<td>1.1</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>19</td>
<td>5</td>
<td>3.00</td>
<td>70.57</td>
<td>25.35</td>
<td>515</td>
<td>11.5</td>
<td>53.1</td>
<td>11.14</td>
<td>1.00</td>
<td>133</td>
<td>18.4</td>
<td>12.1</td>
<td>3.4</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>20</td>
<td>13</td>
<td>11.81</td>
<td>73.94</td>
<td>30.46</td>
<td>47</td>
<td>61.5</td>
<td>98.1</td>
<td>1.58</td>
<td>0.77</td>
<td>176</td>
<td>24.9</td>
<td>108.1</td>
<td>3.7</td>
<td>–</td>
<td>8</td>
</tr>
</tbody>
</table>

Notes. For each NEST, we list the number of stars it contains (n∗), its total stellar mass (m∗), the equatorial coordinates of its center (αB, δB), its average stellar density (ρ̅), its mean radius as computed from the area of the associated convex hull (RH), the semi-major axis, aspect ratio, eccentricity and position angle of the associated ellipsoid hull (a, a/b, e and PA), the mean 1-NNS distance inside the NEST (rN), the separability and dilution factors (αN and βN), the filament it is associated with (Fil) and the loose group number it belongs to as defined in (Kirk & Myers 2011; GKM).

Fig. 3. Spatial distribution of the 20 NESTs identified in this study. The barycenter of each NEST is indicated by a colored plus mark and all the members are colored the same way. Star not pertaining to any NEST are shown as black dots. The MST defined by the NEST is shown with solid gray lines. NEST are numbered based on their increasing right ascension order (see Table 1). The Planck dust emission (217 GHz) is mapped as the background.
Fig. 4. Dilution factor $\beta_N$ as a function of the inter-NEST separation ratio $\alpha_N$. The size of the marks is indicative of the mean first nearest neighbor $r_N$. These ratios are markers of the reliability of the NESTs. Values higher than 2 (dashed black lines) indicate that a NEST is well separated both from the other NESTs and from the stellar population outside the NESTs.

only marginally separated. Increasing $\epsilon$ by 50% did not further alter the set of identified NESTs as no other pair of NESTs are similarly poorly separated. However, decreasing $\epsilon$ by 50%, down to $\sim 30$ kAU ($\sim 0.15$ pc), has resulted in only eight NESTs being identified which corresponds to the most compact “cores” of the fiducial NESTs. The confidence level for these eight detections above random is extremely high due to their high surface density, reaching 99.997%.

In conclusion, while the exact number and detailed properties of NESTs are dependent on the parameters $N_{\text{min}}$ and $\epsilon$, the detection of stellar NESTs in Taurus, as well as their gross properties, are robust results. In the following analysis, we have retained the 20 fiducial NESTs identified in Sect. 3.1. These are embedded in a more dispersed or hierarchically structured population on larger scales.

3.4. NESTs in the context of previously identified groups in Taurus

The identification of loose stellar groups in Taurus is a topic that has a rich history (e.g., Gomez et al. 1993; Kirk & Myers 2011). It is natural to compare the newly identified NESTs to these groups. In Table 1, we indicate to which group identified by Gomez et al. (1993), Kirk & Myers (2011) each NEST belongs. Five NESTs are located at the center of loose groups identified by Kirk & Myers (2011) in a one-to-one correspondence. On the other hand, the remaining 3 loose groups of Kirk & Myers (2011) are in fact substructured, with two or three NESTs in each. Furthermore, our analysis has revealed seven new substructures, each containing between 4 and 6 stars. These substructures could not have been identified by Kirk & Myers (2011), as these authors have used an arbitrary threshold of 10 stars per group. Nonetheless, based on our analysis, we believe that these small NESTs are physical coherent structures, given our 99.85% significance level above random spatial fluctuations.

Fig. 5. Robustness tests of the detection of NESTs in the three main filaments of Taurus. The fiducial NESTs are indicated by black plus symbols and labeled in each panel. Each panel displays alternate sets of NESTs resulting from varying the selecting criteria. Top panel: the NESTs identified if $N_{\text{min}}$ is reduced to three (large red filled squares) or increased to five (small white filled squares). Middle and bottom panels: $\epsilon$ has been increased and decreased by 10% and 50% respectively. In both panels, small orange squares and large dark blue squares represents the results of increasing and decreasing $\epsilon$, respectively.
4. NESTs properties

4.1. NESTs stellar content

The NESTs contain nearly half of the Taurus stellar population, yet their total projected surface area is a small fraction of the star-forming region. Focusing first on the three central main filaments within the cloud, which appear as the main sites of ongoing star formation in Taurus, the projected area of the convex hull formed by all stars they contain is 33 deg$^2$ while the total projected area of all NESTs located within the three main filaments is only 0.27 deg$^2$. Thus, the NESTs cover less than 1% of the projected area of the central filaments. Expanding this analysis to the whole cloud, including the northern and southern stellar components illustrated in Fig. 3, shows that the NESTs cover less than 0.1% of the cloud’s projected area (0.32 deg$^2$ out of 202 deg$^2$).

Consequently, half of the stellar population in Taurus is concentrated in tiny, high density pockets of stars, with stellar densities that range from 50 to 2500 stars pc$^{-2}$, with a median density that is $\approx 100$ times higher than the average density in the whole cloud. This range of stellar surface densities places all NESTs above the median surface densities in nearby star-forming regions (Bressert et al. 2010), including Ophiuchus and Orion. Indeed, the densest NESTs are close to the maximum found in these star forming regions, although it must be noted that their analysis is unable to probe densities exceeding $\approx 1000$ stars pc$^{-2}$.
4.2. NESTs spatial distribution and geometry

To further study the nature of NESTs, we have computed their “ellipsoid hull” which are the ellipsoid of minimal area such that all given stars within the NEST lie just inside or on the boundary of the ellipsoid. These ellipses are characterized by their semi-major and semi-minor axis \((a, b)\), eccentricity \((e)\), centroid \((\alpha_C, \beta_C)\) and major axis position angle \((PA)\). These quantities are listed in Table 1 and allows other studies of NESTs, such as the cumulative distribution of their position angles, the spatial distribution of their centroids, and derivation of their most probable intrinsic 3D geometrical properties; all 3 topics that will be described in the following.

As is evident in Fig. 6, all but two of the NESTs appear tightly concentrated along the three main filaments of the Taurus molecular cloud. Considering both the location and elongated geometry of the NESTs, we have noticed in Fig. 6 that the NESTs appear preferentially oriented along the gas filaments. Beyond this visual inspection, we wanted to go further and present a quantitative argument. To study the relative alignment of the NESTs, we have used the orientation of the local magnetic field traced by linear polarization measurement of background stars as reference. Based on observations, gas filaments in molecular cloud run rather perpendicular to that local direction (Chapman et al. 2011; Planck Collaboration Int. XXXII 2016) even if in denser environments, the orientation may be either parallel or perpendicular (Li et al. 2013; Zhang et al. 2014), although a much more complicated picture has been recently obtained in massive star forming regions (Koch et al. 2018). These findings may be understood in the framework of a recent theoretical work showing that these two configurations (i.e., at low gas column density, magnetic field tends to be orthogonal to the density gradients, while it tends to be parallel to them at high gas column density), are shown to be the two preferred modes that a turbulent magnetized gas found at equilibrium (Soler & Hennebelle 2017). Taurus is well known to be a low gas column density environment, and as such most of the magnetic fields are perpendicular to the main filaments, probably fed by the gas along the striations (Palmeirim et al. 2013). Figure 7 reveals quantitatively that the position angle of the NESTs is indeed preferentially oriented perpendicular to the local magnetic field, along the filaments. In comparison, Ménard & Duchêne (2004) have shown that the symmetry axis of individual YSO disks in Taurus is randomly oriented relative to the magnetic field, while the major axis of dense cores is intermediate between these two populations, neither completely random nor almost always perpendicular to the magnetic field (Lee & Myers 1999; Ménard & Duchêne 2004). This study shows quantitatively the very close connection of the NESTs and the gas filament.

We have then studied the spacings between NESTs by themselves using the MST built from the set of the NESTs centroids (see Fig. 3). We have found that the median length of a segment of the MST is 2.3 pc. However, given the distribution of NESTs along the filaments, the distribution of segment lengths is skewed by those that connect NESTs located in different filaments. Nonetheless, the median length of the segments in the northern two filaments is 1.9–2.1 pc (with a dispersion of \(\approx0.8\) pc), i.e., not significantly smaller. On the other hand, the three NESTs identified in filament 3 are separated by 0.7 and 1.1 pc, respectively, significantly closer to one another. If the NESTs 10 and 12, which are located in that southernmost filament, instead constitute a single, larger NEST (see Sect. 3.3), the NESTs are distributed rather evenly along filaments, with a typical spacing of \(\approx2\) pc.

We now move on analyzing the intrinsic geometry of the NESTs. The apparent elongation of the NESTs informs their three dimensional structure, but projection effects must be taken into account to infer the latter. The simplest three dimensional ellipsoids shapes, such that two of three axis lengths are equal, are either prolate (one major axis \(l_m\), two same minor ones \(l_s\)) or oblate (two same major axis \(l_m\), one minor one \(l_o\)). The aspect ratio \(q\) is then defined as \(q = l_o/l_m\). If one considers an ensemble of spherical spheroids that are randomly oriented in 3D space, the relationship between the intrinsic 3D aspect ratio \(q\) of these spheroids and the expected value of their projected 2D aspect ratio \(q_p\) is given by (Myers et al. 1991) as:

\[
q_p = \left\{ \begin{array}{ll}
\frac{1}{2} \left( 1 + \frac{q}{\sqrt{1-q^2}} \right) \ln \left( \frac{\sqrt{1-q^2} + q}{q} \right) & \text{(oblate case)} \\
\frac{1}{\arccos q} & \text{(prolate case)}
\end{array} \right.
\]

The expected cumulative function of the projected aspect ratio \(q_p\) (Eq. (3)) for randomly oriented 3D oblate and prolate ellipsoids are plotted in Fig. 8 along with the empirical cumulative distribution of the projected aspect ratio \(q_p\) of all 20 NESTs. It is extremely rare for oblate ellipsoids to project into a high aspect ratio structure, as this can only happen if they are observed exactly perpendicularly to their main plane. Therefore, the projected ratio \(q_p\) associated to a distribution of oblate ellipsoids spans the 0.5–1 range, excluding an oblateness hypothesis for the NESTs. The NESTs are then most probably prolate. From the Fig. 8, we see indeed that at small values of \(q_p\), less than 0.35 (i.e., \(q_p<0.25\), highly elongated structures), the observed blue points associated with the NESTs follow the prolate curve. But at higher values, it deviates from a prolate randomly oriented distribution, with half of the NESTs having a value of \(q_p\) around 0.4, its maximal value being 0.7 (\(q_p\) around 0.6). This results suggest that the 3D distribution of the prolate structures are not randomly oriented. It is what we expect for the NESTs structures being aligned with the gas filaments. It is thus tempting to assume that the NESTs are forming inside and along the gas filaments.
filaments, keeping through time their pristine protolate structure despite the dynamical effects (gas expulsion, and dynamical star interaction).

4.3. Mass and size distribution of NESTs

We now study the distribution of the NESTs mass based on sum of the mass of the objects as given in full catalog of Paper I. It should be noted here that the reported mass of a NEST is evaluated by the sum of the primary masses of individual object, as given in Paper I, as masses for stellar companions are incomplete at this stage. It is then a lower mass estimation for the mass of the NESTs. The distribution of the NESTs’ mass is shown in Fig. 9 as $\Phi_N = dN/d\log m$, where $N$ is the total number of the NESTs per logarithmic mass bin and $m$ the mass of the NESTs. It is broad and heavy-tailed, ranging from $\lesssim 1 M_\odot$ to $\gtrsim 10 M_\odot$. In the logarithm mass space, a normal distribution provides an acceptable fit with a mean and standard deviation of $\log m = 0.54 \pm 0.07$ and $\sigma(\log m) = 0.33 \pm 0.05$, respectively. Conversely, the high-mass end ($m \gtrsim 2 M_\odot$) of the distribution is well reproduced by a power law ($\log \Phi_N \propto m^{\gamma}$) of index $\Gamma = -0.50 \pm 0.1$. We note that this index is smaller than the Salpeter reference slope ($\Gamma_S = -1.35$) observed in stellar mass functions (Bastian et al. 2010), but in reasonable agreement with the slope reported for CO clumps (e.g., $\Gamma = -0.65$ and $-0.85$; Kramer et al. 1996; Heithausen et al. 1998) and more recently for massive clumps ($\Gamma = -0.32$, albeit at higher masses than NESTs; Liu et al. 2012). This analysis and this connection made with the gas clump should however be taken with caution due to the low numbers statistics, as indicated by the fairly large uncertainties reported in Fig. 9.

Contrary to the mass distribution, the size distribution of NESTs is clearly bimodal (see left panel of Fig. 10). The two peaks occur at $\approx 12.5$ kAU and $\approx 50$ kAU. Moreover, we have found that the number of YSOs inside each NEST is dependent on the NEST size, possibly revealing two distinct regimes, as illustrated in Fig. 11. All the smallest NESTs ($\lesssim 30$ kAU) contain 4–6 systems, whereas larger NESTs are characterized by a steady increase in the number of members with the NEST’s size.

While we have found no correlation between the NESTs’ radius and their stellar density, there is a positive correlation with their total stellar mass (see Fig. 12). A Pearson correlation test indicates that this correlation is highly significant ($10^{-4}$ false alarm probability). To quantify this correlation, we perform a power law ($m \propto r^\gamma$) Deming fit that takes into account uncertainties on both independent quantities. Uncertainties on the radius of the NESTs are taken to be the difference between the radius of the convex hull ($R_{\text{Hull}}$) and the geometrical radius computed from the minimum spanning ellipsoid fit ($R_G = \frac{1}{\sqrt{\pi}}ab$). The uncertainty on the total mass is dominated by uncertainties on the individual stellar masses, which may be as high as 30% (Kirk & Myers 2011). We adopt this conservative estimate in our analysis. From the Deming fit, we derive $\gamma = 0.94 \pm 0.21$, i.e., a nearly linear correlation. Thus the mass-radius relationship for the NESTs is markedly shallower than that of star clusters (for which the power law exponent is about 5/3, Pfalzner et al. 2016, and references therein) and of prestellar clumps and cores in the simulations (Lee & Hennebelle 2016, and references therein). On the other hand, the nearly linear behavior is reminiscent of the correlation expected for isothermal critical Bonnor–Ebert spheres, the assumed conditions setting the onset of gravitational collapse against thermal support. In that situation, the power law exponent is predicted to be exactly unity. This is consistent with observed core properties, as summarized in the work of Motte et al. (2018, see their Fig. 7). As initially reported in Motte et al. (2001) and confirmed by Könyves et al. (2015), the prestellar cores that are dominated by gravity are close to the linear relation $m \propto r$, unlike the cores dominated by turbulence which have the power law $m \propto r^2$. 

![Fig. 8.](image_url) Cumulative distribution of the projected axial ratio $q_p$ (Eq. (3)) for a population of randomly oriented oblate (black) and prolate (red) ellipsoids for a fixed intrinsic (3D) axial ratio $q$. The blue solid circles mark the cumulative distribution of the 2D ratio $q_p$ of the NESTs.

![Fig. 9.](image_url) Mass Function of NESTs (dashed histogram). The graphic shows the mass probability density function ($\Phi_N = dN/d\log m$) as a function of $\log m$, where $N$ is the total number of the NESTs and $m$ the mass of the NESTs. For comparison, we report also the mass functions of the H$^{13}$CO$^+$ cores (Onishi et al. 2002, light purple) and the near-infrared extinction dust cores (Schmalzl et al. 2010, light green). The red curve and black dashed curve indicate respectively the power law $dN/d\log m \propto m^{-1}$ fit for the high mass NESTs distribution (red points) and normal fits to the whole observed mass distribution expressed in logarithm of the mass.
Fig. 10. Size distribution of NESTs (which exhibits two peaks, at 12.5 and 50 kAU, respectively; left panel), H$_{13}$CO$^+$ cores (Onishi et al. 2002; central panel) and near-infrared extinction dust cores (Schmalzl et al. 2010; right panel).

Fig. 11. Number of YSOs in NESTs as a function of their size.

5. Discussion

5.1. NESTs: Preferred sites of star formation

Comparing the population of YSOs located inside and outside of NESTs can give us information on the physical nature and origin of NESTs. Most importantly, Class I sources are the most embedded, hence likely the youngest, YSOs in Taurus. Sites of ongoing star formation are therefore expected to host a high fraction of Class I sources. Table 2 summarizes the classification of all objects inside and outside of NESTs.

The relative proportions of Class II and, III sources inside the NESTs is not significantly different from that of the more
dispersed population, despite a slightly lower proportion of Class III sources. On the other hand, the NESTs host a proportion of Class I sources in Taurus that is ≈3 times higher than the stellar population outside of NESTs (18.5 and 5.9%, respectively). A standard Pearson χ² statistical test indicates that this difference is significant at the 99.9% confidence level, in other words Class I sources are preferentially found within NESTs.

Besides the fact that almost 75% of all Taurus Class I sources belong to NESTs, they are concentrated in just 11 of those. This high concentration reveals that these NESTs represent regions the most presently active in Taurus. The remaining nine NESTs contain only Class II and III objects suggesting that nearly half of the NESTs are getting infertile while the other half has experienced recent star formation. Indeed, the ratio of the number of Class I versus Class II sources and the ratio of Class II to Class III sources within NESTs suggest an evolutionary temporal scheme (see Fig. 13). In particular, the proportion of Class I objects decreases as the proportion of Class III objects increases. When the fraction of Class III objects inside a NEST reaches 60%, there are no associated Class I objects, suggesting that these NESTs are the oldest. Conversely, NESTs with a high proportion of Class I objects (80%) have no Class III objects. A NEST with such a YSOs content is thus amongst the youngest.

Past studies in Taurus (e.g., Luhman et al. 2010) show that Class III sources are more dispersed than Class II themselves being more dispersed than Class I objects. In our catalog, the mean 1-NNS of all Class I, II and III objects are 0.15, 0.34 and 0.51 pc, respectively (see Table 3). This has been interpreted in the past as a consequence of the least evolved objects being preferentially concentrated near the gas filaments and the more evolved ones being much more widely dispersed.

The identification of NESTs as physical substructures within Taurus prompts us to revisit this finding. For all classes, objects located inside the NESTs have larger mean 1-NNS than the overall population and these mean 1-NNSs show a marked increase from Class I to Class III. This is in line with the previous interpretation of a dispersed population of Class III sources throughout the cloud. However, the mean 1-NNS for objects located within NESTs shows a qualitatively different behavior. Strikingly, while there is also a slight marginal increase of the 1-NNS from Class I to Class III objects, all these values are consistent with a single value for all the 3 Classes given their associated uncertainties (see Table 3). This shows that, inside the NESTs, neither Class II nor Class III sources are more widely spread out than Class I sources, suggesting that the spacing between objects of a given class within NESTs does not depend on evolutionary stage.

In an effort to search for signs of the expansion of NESTs, we searched for any correlation between the population and proportion of Class I, II and, III and their radius or surface density but found none. This suggests that there is no detectable dynamical expansion of NESTs at this scale, i.e., NESTs remain tightly packed as they age. In that respect, the observed NESTs may be pristine spatial imprints of the stellar distribution at birth. Most interestingly, the spatial distribution of the lowest mass YSOs appears different at a significant level than that of the most massive ones. Specifically, we have found that the majority of the former are found outside the NESTs, in the more distributed population. This aspect will be developed in more details in a forthcoming paper.

### 5.2. NESTs as intermediate spatial scale structures

With a typical size of 0.1–0.25 pc, the NESTs are smaller structures than the loose groups identified in the past in Taurus (e.g., Kirk & Myers 2011) and are more comparable to the size of the largest UWPs, which contain more than half of all sources in Taurus. By construction, traditional multiple systems are much smaller than NESTs. In Paper I, we have shown that the majority of UWPs have high order multiplicity, especially the closest ones with separation in the 1–10 kAU range. Characterizing the connection between structures on these different spatial scales is important to assess whether they arise from a unique process, whereby NESTs and UWPs could be pristine imprints of a fragmentation cascade scenario.

To probe how these different scales of stellar groups are related, from the multiple scale (5–1000 AU) to the UWPs scale (1–60 kAU) up to the NESTs scale (15–155 kAU), we estimated the fraction of YSOs that belong in each stellar group of each NESTs.
The vast majority of YSOs are either members of multiples, UWPs or NESTs, leaving a small fraction less than 20% of single isolated stars. Only 20% of multiples are not included in groups of higher hierarchy (i.e., either within UWPs or NESTs) whereas nearly half of the UWPs are inside NESTs. Only ~20% of the whole YSO population (single and multiples) belong neither to an UWP nor to a NEST. They are isolated since they are located at the peripheries of the stellar groups and on average 4.7 times further away (0.56 pc, 115 kAU for the median) from their first nearest companion compared to the other YSOs that belong either to UWPs or NESTs (0.12 pc or 25 kAU for the median). Although this isolated population contains proportionally more Class III objects and less Class II and I objects, this difference does not appear statistically significant when performing the Pearson’s contingency test.

We propose to outline the connection between the different length scales through the estimate of their surface stellar density as a function of their mean radius $r$. For UWPs, the latter is estimated as half their separation, and the stellar surface density $\rho_\star$ is then estimated as the total number of stars $n_\star$ within the UWPs divided by the area, i.e., $\rho_\star = n_\star/(\pi r^2)$. For NESTs, loose groups (as defined in Kirk & Myers 2011), the three main filaments region and the whole Taurus cloud, the projected area is estimated from the surface $A$ of their associated convex hull containing each $n_\star$ stars. We fit an ellipsoid to the convex hull and derive the semi major and minor axes (respectively $a$ and $b$). We have defined the mean geometrical radius $r$ as $r = \sqrt{ab}$ and the associated stellar surface density $\rho_\star = n_\star/A$.

The surface stellar density of these spatial features, shown in Fig. 15, reveals two regimes: one is associated with UWPs ($\rho_\star \propto r^{-2}$), and the second with loose groups and spatial features on larger scales ($\rho_\star \propto r^{-1.2}$). The former is related to cascade fragmentation (Paper I), while the latter is associated with a clustering regime. NESTs appear as intermediate structures, with smaller (resp. larger) ones following the relationship observed for UWPs (resp. loose groups and larger structures).

5.3. Origin of NESTs

The close connection between the NESTs and the gas along the Taurus filaments raises the question of their origin (see Fig. 6). We have distinguished, as generally accepted, the more dense cores with typical density of $10^{18} - 10^{20}$ cm$^{-3}$ and length scale of 0.05–0.1 pc from larger diffuse gas clumps with typical density of $10^{-3} - 10^0$ cm$^{-3}$. Zooming in the L1495 Taurus region and gathering data from NIR dust extinction and C$^{18}$O clumps (Schmalzl et al. 2010; Hacar et al. 2013) and dense N$_2$H$^+$ and H$_{2}$CO$^+$ cores (Hacar et al. 2013; Onishi et al. 2002), we indeed note that the large scale structure of filament as outlined by the C$^{18}$O gas tracer is closely associated with the location of both NESTs and molecular cores (see Fig. 16). We note also that there are no NEST in regions where dense cold molecular cores are the most present and clustered (B216, B218, B210, B10). The exceptions are at the edges of both N$_2$H$^+$ B213 clustered cores regions (NESTs numbers 3 and 4) and for the largest NEST (NEST number 2) located in the B7 region, where there are few dense N$_2$H$^+$ cores in the central part. These three situations may illustrate different evolutionary states for the formation of stars in those regions, the latter being the most evolved.

The range of surface density in NESTs is close to that of dense cores in molecular clouds with typical H$_2$ surface density of $10^{20} - 10^{22}$ g cm$^{-2}$ (see Fig. 12). Moreover, we have gathered mass information on NIR dust extinction clumps (Schmalzl et al. 2010) and H$^{13}$CO$^+$ cores (Onishi et al. 2002). While a KS test indicates that the two associated core mass distributions differ from one another (at the 99.8% confidence level), the mass distribution of NESTs is indistinguishable with either ($p$-value of the same parent distribution of 0.8 and 0.1 for NIR or H$^{13}$CO$^+$ cores, respectively).
These clues tend to suggest that NESTs are the direct descendents of cores, in the sense that each NEST could be connected on a one-to-one basis to a dense core that subsequently collapses and fragments. Indeed, observations showed multiple and wide pair Class 0 or 1 objects within a single core (Pineda et al. 2015; Sadavoy & Stahler 2017), indicating that several objects may form within a single core. However, while this relation is plausible for the small NESTs having 4–6 stars, this scenario appears less credible to explain the formation of the richer NESTs sheltering up to more than 20 stars from one single core. Indeed, to form these richer NESTs would require denser cores. Since such cores are not observed in Taurus (Hacar et al. 2013; Onishi et al. 2002), a scenario implying aggregation of cores is necessary to provide an explanation for the richer NESTs.

The observed bimodality in the distribution of NEST sizes and the two regimes associated with their size-star number relation suggest there are two mechanisms at work in the formation of NESTs, allowing us to produce a coherent framework for all NESTs. The first peak in the distribution of size appears in the range of less dense gas cores as traced by near-infrared extinction maps (Schmalzl et al. 2010). Besides, the mass distribution of the NESTs spans the same range of both mass cores.

It is therefore tempting to associate the first regime of NESTs (those with few members) with the fragmentation of a single core and the second regime (the richer ones) with a clustering of a few cores in a chain, as it is outlined in the work of Tafalla & Hacar (2015). Indeed these authors report that the distances to the nearest neighbor among the N2H+ cores in L1495 mostly lie below 0.2 pc, 41 kAU, exactly the range of the second peak. They also find that these dense cores tend to cluster in linear groups of three cores on average, which they call chains. The elongated geometry of the NESTs would also be compatible with this scenario, either due to the probable nature of the cores (Myers et al. 1991) or the linear fragmentation of the filament given the chain of cores as observed by Tafalla & Hacar (2015). The process of hierarchical fragmentation from the cloud to protostars has been recently outlined in the Perseus complex based on different gas tracers, in particular different paths for cloud fragmentation (isolated cores and cores in filaments, Pokhrel et al. 2018). We propose that the present study outlines the stellar structures resulting from this hierarchical fragmentation. Viewing the NESTs as the intermediate spatial structures between the UWPs and the loose groups, part of the UWPs should also result from this clustering scenario of cores as proposed by Tokovinin (2017). Furthermore, the power law fit at the high mass range of the NESTs $\Gamma = -0.5$ is much shallower than the power law found for dense cores in Taurus, as Sadavoy et al. (2010); respectively Schmalzl et al. 2010) found $\Gamma = -1.22$ (resp. $\Gamma = -1.2$) for the starless dense cores (resp. NIR extinction clumps) more massive than $2M_\odot$. But close to the the power law associated with larger CO clumps ($\Gamma = -0.65$ (resp. $\Gamma = -0.85$), Kramer et al. (1996);
5.4. Links to star formation models

Young stars form on a wide range of scales out of their molecular parental cloud, producing aggregates, groups, clusters and distributed population with various degrees of clumpiness and stellar density. But whether these star forming regions form as a result of a slow collapse or contraction scenarios of clumps at quasi-equilibrium over several free-fall dynamical times (e.g., Krumholz et al. 2006; Tan et al. 2006; Farias et al. 2017) or they collapse more quickly (intermediate or short timescales) on the order of 1–2 dynamical time (e.g., Ballesteros-Paredes et al. 1999; Elmegreen et al. 2000; Hartmann et al. 2001; Bate et al. 2003) mediated either by large-scale accretion flows along filaments (Myers 2009; Smulys & Bonnell 2017) and/or a global collapse (Hartmann & Burkert 2007; Vazquez-Semadeni et al. 2017) remains an open question.

The possible diagnostics to discriminate between these models include the morphologies of subclusters, the degree of clustering and hierarchy, and the spatial distribution of stars. We intend to develop in our next paper in this series the analysis of the subclusters hierarchy in Taurus to expand in that direction. For now, focusing on the densest part of Taurus, i.e., NESTs, may already prove informative. In the slow type models, the morphology of subclusters tend to be rather spherically symmetric (Tan et al. 2006) in contrast with the rapid type model in which elongated structures dominate (Bate et al. 2003). Thus, the highly elongated NESTs we found tend to favor the rapid scenario. This is in agreement with Hartmann et al. (2001), who have made the assumption of a rapid formation and dissipation of molecular clouds in a few dynamical times to explain the small age spread (1–3 Myr) of the Taurus members and the absence of the Post T Tauri stars. However in all scenarios, the most recent works show that the subclusters that form at local scale merge afterwards within larger ones that themselves merge to finally join a large central cluster fueled with gas and stars. But if the mixed stellar populations we observed within NESTs were to be due to the merging of previous physically unrelated sub-units, we do not expect to see an evolution of the status of the population inside the NESTS as we have noted. An important caveat is that most of these models, and specially the slow type models, have been developed to simulate massive and dense star forming regions, and cannot be directly compared to the Taurus region. To our knowledge, no specific model to explain the formation and the persistence of the NESTs in Taurus has been proposed.

To conclude this section, we suggest that the NESTs are pristine imprints of stellar formation and that they are representative of two fragmentation scenarios. The first one is associated with the fragmentation of a single dense core that could give birth to a small number of stellar systems (4–6), while the second one corresponds to the fragmentation of a filament in a few (2–5) cores closely spatially associated, and thus spatially clustered. These structures may remain mostly unchanged for at least a few millions years.

6. Conclusion

Building on the stellar spatial distribution in the Taurus star forming region, we have used the density based clustering algorithm DBSCAN to identify local overdense subclusters. We have set the two free parameters of the algorithm based on the one-point correlation function and the cumulative function of the k-nearest neighbor separations, in order to reach a 99.8% significance level of detection above random. We found 20 stellar structures that we dub NESTs (Nested Elementary STructures). We have then performed a set of statistical studies to derive their main properties.

Our work has shows that about 45% of the whole stellar population in Taurus is concentrated in 20 small and most probably prolate (median eccentricity of 0.9) NESTs that are regularly spaced ($\approx 2$ pc) and mainly oriented along the principal gas filaments axes. Five of these NESTs are associated with the densest parts of eight previously identified loose groups (Gomez et al. 1993; Kirk & Myers 2011). Each of the 3 remaining loose groups are composed by two or three stellar NESTs and seven of the NESTs were not identified so far. The stellar NESTs contain between 4 and 23 stars each, the median being 5–6. Inside the NESTs, the surface density of stars may be as high as 2500 pc$^{-2}$ (while the mean surface density of the Orion Nebula Cluster is about 1,000 YSOS/pc$^2$. Bressert et al. 2010). The mean value is of the order of 340 pc$^{-2}$, which agrees well with the initial substructures density (350 pc$^{-2}$) required to reproduce the present-day binary properties in Taurus when starting from a universal initial binary distribution according to Nbody numerical simulations (Marks & Kroupa 2012).

Although the proportion of Class II and, Class III objects inside the NESTs is not significantly different from that of the more dispersed population, the NESTs as a whole contain 3/4 of the class I objects. Studying them individually, only half (11) of these NESTs contain those class 0 or I objects, showing that they are the sites of privileged star formation. The balance between Class I, II and, III fraction within the NESTs suggests that they may be ordered as an evolutionary temporal scheme, some of them getting infertile with time, while other still giving birth to young stars. The fact that Class III objects are still present in such a relatively tight environment suggests that they may get old and remain in the same environment for few millions years. Furthermore, the higher mass stars of the population are equally found inside and outside the NESTs, but the great majority (60–80%) of their lowest mass counterparts are found outside the NESTs. This specific point will be developed in a forthcoming paper. The total mass of stellar NESTS ranges from half a solar mass up to 10 $M_\odot$, with a median mass of 3 $M_\odot$. The size distribution of the NESTs is bimodal with one peak at 12.5 kAU and the second at 50 kAU.

We have identified the preferred sites of star formation in Taurus as the densest stellar groups of the region. Each NEST may be the individual stellar outcome of the gravitational collapse of a cloud that fragments to give groups of stars within few millions years. Cloud fragmentation being prior to the formation of the YSOs, it could lead to the observable clustering of dense cores and then stars. These NESTs are intermediate structures between the ultra-wide pairs and the loose groups, and provide an explanation for the elbow observed in the two-point correlation function and mark the transition between the multiplicity regime and larger scale structures. The study of the relationship between the different scales of stellar groups, NESTs, UWPs and multiple systems, reveals that they are inter-connected, with only 20% of stars truly single and isolated.

We propose that the youngest NESTs are the spatial imprints of the stellar distribution as they may have emerged from their natal cloud at a scale that have been overlooked up to now in Taurus. Using DR2 GAIA release should provide invaluable information both on distance to probe the 3D structure and on kinematics to further probe the dynamical status of these features and their origin. We also intend in future work to analyze other
Appendix A: Catalogs of stars in NESTs

Table A.1. Members of Taurus belonging to NESTs

<table>
<thead>
<tr>
<th>#</th>
<th>2MASS</th>
<th>Name</th>
<th>RA_J2000</th>
<th>DEC_J2000</th>
<th>SpT</th>
<th>M (M⊙)</th>
<th>Class</th>
<th>n,</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>J04135328+2811233</td>
<td>IRAS04108+2803A</td>
<td>63.472</td>
<td>28.19</td>
<td>M4c</td>
<td>0.271</td>
<td>I</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>J04135471+2811328</td>
<td>IRAS04108+2803B</td>
<td>63.478</td>
<td>28.19</td>
<td>M2c</td>
<td>0.575</td>
<td>I</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>J04141188+2811535</td>
<td>-</td>
<td>63.550</td>
<td>28.20</td>
<td>M6.25</td>
<td>0.086</td>
<td>II</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>-</td>
<td>IRAS04111+2800G</td>
<td>63.551</td>
<td>28.14</td>
<td>M2c</td>
<td>0.575</td>
<td>I</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>J04141291+2812124</td>
<td>V773TauA+B</td>
<td>63.554</td>
<td>28.20</td>
<td>K3</td>
<td>1.796</td>
<td>II</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>J04141358+2812492</td>
<td>FMTau</td>
<td>63.557</td>
<td>28.21</td>
<td>M0</td>
<td>0.701</td>
<td>II</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>J04141700+2810578</td>
<td>CWTau</td>
<td>63.571</td>
<td>28.18</td>
<td>K3</td>
<td>1.796</td>
<td>II</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>J04141760+2806096</td>
<td>CIDAI</td>
<td>63.573</td>
<td>28.10</td>
<td>M5.5</td>
<td>0.137</td>
<td>I</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>J04142639+2805597</td>
<td>MHO2</td>
<td>63.610</td>
<td>28.10</td>
<td>M2.5</td>
<td>0.486</td>
<td>II</td>
<td>3</td>
</tr>
<tr>
<td>10</td>
<td>J04143054+2805147</td>
<td>MHO3</td>
<td>63.627</td>
<td>28.09</td>
<td>K7</td>
<td>0.801</td>
<td>II</td>
<td>2</td>
</tr>
<tr>
<td>11</td>
<td>J04144739+2803055</td>
<td>XEST20-066</td>
<td>63.697</td>
<td>28.05</td>
<td>M5.25</td>
<td>0.157</td>
<td>III</td>
<td>1</td>
</tr>
</tbody>
</table>

Notes. This table was built from the catalog in Paper I Columns 1–3: Star ID in the table, Identification number of the NEST 2MASS Point Source Catalog and common Name. Column 4–5: Ecliptic right ascension and Declination (Epoch J2000). Columns 6–8: Spectral type, (Primary) mass, Class. Columns 9: Total number of stars (n,) within 1000 au.

Table A.1 lists all Taurus members that are located within NESTs. Details on the multiplicity of each systems (separations and references) can be found in Table C.1 of Paper I.

Appendix B: DBSCAN algorithm

In this appendix, we define more formally the concepts and vocabulary of dbscan (Density-Based Spatial Clustering Applications with Noise) clustering method, along with complementary comments relevant to our analysis. Some terms are taken from the work of Ester et al. (1996).

The n objects of the dataset are assumed to be embedded in a d-dimensional space \( \mathcal{R} \subset \mathbb{R}^d \). In our analysis of stellar clustering, we will deal with a projected Euclidean space of dimension \( d = 2 \).

### B.1. Basic concepts of DBSCAN

The \( \epsilon \)-neighborhood of an object \( i \) in \( \mathcal{R}^d \) is defined as the subset \( [\epsilon_i] \) of \( \mathcal{R}^d \) such that:

\[
[\epsilon_i] = \{ j \in \mathcal{R}^d \mid d(i, j) \leq \epsilon \} \quad \epsilon \text{-neighborhood of } i,
\]

(B.1)

where \( d(i, j) \) is the distance between objects \( i \) and \( j \). We now introduce two intermediate definitions. First, two objects \( i \) and \( j \) of the dataset are said to be \( \epsilon \) direct-neighbors, if their distance \( d(i, j) \) is less or equal to \( \epsilon \), in other words, if \( j \) belongs to the \( \epsilon \)-neighborhood of \( i \) and vice-versa:

\[
\begin{align*}
& i \in \epsilon_i \\
& j \in \epsilon_j
\end{align*}
\]

(B.2)

\( \epsilon \)-direct-neighbors condition.

Secondly, two objects \( i \) and \( j \) of the dataset are said to be indirect \( \epsilon \)-neighbors, if we can find a chain \( l_1, \ldots, l_m \) of \( \epsilon \)-direct-neighbors in \( \mathcal{R}^d \) connecting \( i \) and \( j \). So the indirect \( \epsilon \)-neighbors condition reads as:

\[
\exists l_1, \ldots, l_m \in \mathcal{R}^d \text{ such that } l_1 = i \text{ and } l_m = j \\
\forall m = (1, \ldots, n-1), (l_m, l_{m+1}) \text{ are } \epsilon \text{-direct neighbors.}
\]

(B.3)

Describing our dataset in terms of vertices (nodes) associated with the objects linked together if \( d(i, j) \leq \epsilon \), the direct- and indirect \( \epsilon \)-neighbors conditions reflect the definition of a path within the graph framework.

The \( \epsilon \)-neighborhood population \( N^\epsilon_i \) of a point \( i \in \mathcal{R}^d \) is then defined as the number of objects in \( \mathcal{R}^d \) found within the \( \epsilon \)-neighborhood of \( i \):

\[
N^\epsilon_i = \text{Card}(\epsilon_i) \quad \epsilon \text{-neighborhood population of } i,
\]

(B.4)

where Card stands for the cardinality.

To define clusters, a minimal local neighborhood population threshold \( N^\epsilon_{\text{min}} \) is introduced to identify objects of the dataset having at least the same \( \epsilon \)-neighborhood population or higher. An object \( i \) is said to be a core cluster object if its \( \epsilon \)-neighborhood is as populated or even more, than the given local neighborhood threshold:

\[
N^\epsilon_i \geq N^\epsilon_{\text{min}} \quad \text{core object condition.}
\]

(B.5)

An object \( j \) of the dataset is said to be \((\epsilon, N^\epsilon_{\text{min}})\) directly density-reachable from \( i \) if they are \( \epsilon \)-direct-neighbors and if \( i \) is a core object:

\[
\begin{align*}
& i \in \epsilon_i \setminus N^\epsilon_{\text{min}} \quad \text{direct density-reachability property from } i. \\
& \exists j \in \epsilon_i \quad N^\epsilon_j \geq N^\epsilon_{\text{min}}
\end{align*}
\]

(B.6)

Similarly, a point \( j \) is said to be \((\epsilon, N^\epsilon_{\text{min}})\) density-reachable from \( i \), if \( i \) and \( j \) are \( \epsilon \)-indirect-neighbors through a chain \( l_1, \ldots, l_{n-1} \) of core objects. The density-reachability property reads as:

\[
\exists l_1, \ldots, l_n \in \mathcal{R}^d \text{ such that } l_1 = i \text{ and } l_n = j \\
\forall m = (1, \ldots, n-1), (l_m, l_{m+1}) \text{ are } \epsilon \text{-direct-neighbors,} \\
\forall m = (1, \ldots, n-1) N^\epsilon_{l_m} \geq N^\epsilon_{\text{min}}
\]

(B.7)

Once an arbitrary core cluster object - called the seed - is identified, the density-reachability property allows to retrieve all the objects of the dataset belonging to the same cluster. A cluster is then defined as the set of points that are density-reachable from a seed. If an object \( i \) is density-reachable from a seed \( j \) but does not fulfill the core object condition (Eq. (B.5)), then it is said to be a border cluster object:
Given $\epsilon$ and $N_{r_{\text{min}}}^e$, the \texttt{dbscan} algorithm proceeds in two steps: (1) pick one object from the dataset, check if it is a seed (i.e., satisfying Eq. (B.5)) and if it is not, pick another one until it satisfies the core object condition; and (2) retrieve all the objects in the dataset which are density-reachable from the seed. We note that the core points set in the final clusters obtained at the end of the algorithmic scan does not depend on the chosen seed, since the core objects of a given cluster are density reachable from each other. A cluster $C$ is defined by the following conditions:

$$\begin{align*}
\exists i_{\text{seed}} \in \mathcal{R}^d \text{ such that } i_{\text{seed}} \text{ is a core point of cluster } C \quad &\iff i \in \mathcal{R}^d, \forall i \in (\epsilon, N_{\text{min}}) \text{-density-reachable from } i_{\text{seed}}. \\
\{C\} = \{i \in \mathcal{R}^d, i \text{ is (} \epsilon, N_{\text{min}}) \text{-density-reachable from } i_{\text{seed}}\}. 
\end{align*}$$

(B.9)

If an object of the dataset is not $(\epsilon, N_{r_{\text{min}}}^e)$-density-reachable from any core objects then it is marked as an outlier, or noise, with respect to that property. Noise and outliers are then the set of objects that do not belong to any cluster in the dataset. Based on the above definition and within a graph framework, \texttt{dbscan} clusters are in fact the maximal connected component sets $C$ such that $\forall (i, j) \in C$, $N_i \geq N_{\text{min}}$ and there is a path between $i$ and $j$. This definition of clusters is therefore connected to the more intuitive idea of density-contour clusters outlined by Hartigan (1975).

B.2. Discussion on \texttt{dbscan}

B.2.1. Complexity

In the general case, the complexity of hierarchical clustering techniques is $O(n^2)$, which makes them too slow for large data sets. However, for the single and complete linkage clustering, this is optimally reduced to $O(n^2)$. The \texttt{dbscan} clustering algorithm visits each point of the database possibly multiple times (e.g., as candidates to different clusters). Without the use of an accelerating index structure, the run time complexity is $O(n^2)$. This can be reduced to $O(n \log n)$ with a spatial indexing using a R-tree.

B.2.2. Minimal size of a cluster

There are two free parameters in the \texttt{dbscan} algorithm: $\epsilon$, which defines the size of the neighborhood, or “window,” that is investigated around each object, and $N_{r_{\text{min}}}^e$, the minimal local neighborhood population threshold, which eventually corresponds to a local density threshold since the population number is defined within an $\epsilon$-neighborhood.

Since, $N_{r_{\text{min}}}^e$ is the minimum number of objects that must exist in the $\epsilon$-neighborhood to obtain a core point, it is thus also the minimum size of a cluster (Ester et al. 1996). At least, this minimal cluster is composed by one core point and $(N_{r_{\text{min}}}^e - 1)$ border points. According to the definition of a border point (Eq. (B.8)), it may happen that a point is a border point for several different clusters. Then the labeling of this ambiguous point to a specific cluster depends arbitrarily on the order to which the points are processed during the algorithmic scan. By convention, it is automatically labeled to the first cluster it meets as border point. The algorithm is one of “exclusive” type, as opposed to non-exclusive clustering algorithm for which one point may belong to multiple clusters. If this point appears to be at a later stage also a border point for another cluster and if it is one of the possible smallest cluster, composed then a priori by $N_{r_{\text{min}}}^e - 1$ border points, this cluster would have consequently in fact less members than $N_{r_{\text{min}}}^e$ local population threshold. We note, however, that this fact does not introduce inconsistency with the definition of the density-reachable based cluster $\texttt{dbscan}$ although it highlights the order-dependence of the algorithm. It also outlines two different roles of border points that may vary from being a real frontier that distinguishes a cluster from noise or other clusters from being a bridge between clusters.

B.2.3. Minimum distance between two clusters

The two parameters $N_{r_{\text{min}}}^e$ and the $\epsilon$-neighborhood length play a key role in identifying clusters. If no seed point is found, the whole dataset is considered as noise. Conversely, if a first seed core point is found within the dataset and if all the the points are density-reachable from the seed, we obtain a single cluster containing all objects. The algorithmic process will give rise to a second cluster when only a subset of all points are associated with a core point and a second core point that is not density reachable from the first seed is found in the set. It is then worthwhile to investigate the two conditions at which the points are not density reachable. The first one states that a point is never density reachable from a seed point if it is not its $\epsilon$-indirect neighbor. The second point outlines the fact that a point is never density reachable from a border point, even if this point is in the $\epsilon$-neighborhood of the border point.

So there are two basic situations that stop the ongoing growth of a cluster. One is based on the $\epsilon$-neighborhood length threshold since the algorithm excludes as candidate members all the points that are not $\epsilon$-indirect neighbors of the seed (first condition above). The second is based on $N_{r_{\text{min}}}^e$ parameter which is the parameter that is indirectly used in order to decide whether a cluster should be increased at a certain stage of the algorithm. This parameter describes in fact the minimum number of reachable points that a given point must have in its $\epsilon$-neighborhood not to be a border point. A significant local drop of $\epsilon$-neighborhood density within the dataset therefore stops the growth process (second condition quoted above). In other words, the cluster growth stops when no more path is found from a core point, when the edges of a cluster are surrounded by less dense regions. Either way, the first conditions implies that two clusters are separated by a distance of at least $\epsilon$ to discriminate them from being one single cluster. However, this threshold value is only true when the distance between clusters is estimated from core points. This separability distance may indeed be shorter when border points are implied in this inter-distance evaluation (second condition). When this occurs, the two clusters are marginally separated due to a localized drop in density. Choosing a shorter $\epsilon$-neighborhood length to scan the dataset would lead to identifying these border points as noise and would separate more clearly the two clusters.

B.2.4. Range of the number of star components

Given the $\epsilon$-neighborhood length parameter, the $N$ objects of the dataset are partitioned into a set $\mathcal{O}$ grouping data that are noise or outliers with $0 \leq N_O = \text{Card}(\mathcal{O}) \leq N$ and $N_{C}$ number of clusters. When no seed cluster is found, the condition Eq. (B.5) is never fulfilled, no cluster is detected and $N_O = N$ and $N_C = 0$. Excluding the pathological cases implying “ambiguous points”, the theoretical maximum value of the number of the clusters $N_{C}$ that can be found to partition the dataset is obtained when the neighborhood of each core point is composed by $(N_{r_{\text{min}}}^e - 1)$ border points. The total population of neighborhood is then exactly equals to the minimal value $N_{r_{\text{min}}}^e$. Then, the maximal
range of the number of clusters is given by
\[ 0 \leq N_c \leq E(N/N_{\text{min}}) \]
where \( E \) is the floor function.

**Appendix C: Discussion on clustering algorithms**

This appendix is intended to review the different types of clustering algorithms in order to provide the context for our choice of the one level \texttt{dbscan} algorithm as the optimal procedure to detect local overdose spatial structures.

Clustering is the process of examining a collection of objects embedded in a \( d \)-dimensional space to group closest objects together automatically into natural clusters, while distinguishing groups from each other. Although this task may sometimes be easily done by visual inspection in a 2-dimensional space, it is not straightforward to formalize the process on general grounds. The fundamental challenge is to conceptually and quantitatively define the meaning of close, hence to define from heterogeneous variables a proximity (or distance) function between objects. There is no general law, so at best the criteria used are heuristic, and ultimately will depend on the clustering purpose of the user.

A traditional class of hierarchical clustering algorithms are the so-called agglomerative, or bottom-up, algorithms. These non-parametric algorithms start from initial set of clusters (each composed of a single object) and iteratively build nested families of larger clusters through successive mergers of cluster pairs, using a heuristic proximity function criteria, until one single clusters containing all objects in the distribution is obtained.

The other traditional class of clustering algorithms are the non-parametric divisive hierarchical clustering algorithms, which essentially are the reverse of the agglomerative algorithms. That is to say that such algorithms start from the entire set and, step by step, divide the set in subsets until single objects are obtained. However, since the divisive approach is somewhat more complicated to implement (for a review on spectral partitioning, see Filippone et al. 2008), here we focus on the agglomerative algorithms.

In agglomerative algorithms a similarity (i.e., proximity) or dissimilarity (i.e., distance) function is defined heuristically at each merging step as described by the linkage method being employed, which is selected by the user prior to implementing the algorithm. In this approach, six most widely employed linkage methods are generally used, depending on the distance criteria chosen to proceed to the merging of clusters. Firstly the “single Link” method has a merging criteria based on the distance between the closest single element pair of two clusters in the set. Secondly, the “complete Link” method is based on the distance between the furthest element pair of two clusters in the set. Thirdly, the “average Link” method is based on the distance between each element pair of two clusters in the set. The other two methods, “centroid” and “median” links are respectively based on the distance between the centroids of two clusters and the median distance between the each element pair of two clusters in the set. Although similar to the average link method, the median link method is less sensitive to outliers. Finally, the fourth method is the “Wards’ link based on the value of the sum of squared deviations between the projected centroid of two clusters in the set.

In all these bottom-up methods, the cluster pair in the set with the smallest distance are then merged. The strength of these non-parametric methods is that no assumptions are made about the structure of data. However, they suffer from a significant drawback: since each step of the agglomerative process is built upon the previous one, and no backtracking is permitted, there is a loss of data such that the nested hierarchical structure is built-in and not the result of an open process free of reconfiguring the structure of data at each step. In other words, the hierarchical clustering algorithm is sensitive to possibly erroneous previous cluster merging, since, once assigned, objects’ cluster memberships are not permitted to change (i.e., the assignments at each step are permanent).

Another significant drawback of these agglomerative methods is their lack of robustness, especially when the distribution contains noise and outliers. Generally speaking, this leads spurious additional clusters being identified or the blurring of distinct clusters. It also tends to produce clusters of a particular shape. For example, in the single linkage method, the merging of two clusters depends only on a local merging criteria based on their most proximate members. Thus, only the closest parts of the two clusters are considered whereas the overall structure of the clusters and their more distant parts are not taken into account. Since clusters are merged on the criteria of the closest element pair of two clusters in a set, it may happen that in a noisy distribution elements are successively merged in such a way that two obvious distinct clusters are merged together (due to few sparse singletons being close to each other), even though many of the elements in each cluster are very distant from each other. This tendency to combine elements linked by a series of close intermediate elements gives rise to long chains (this is known as the chaining effect). On the other hand, in the complete linkage methods, the merging of two clusters at one step is based on the proximity of their furthest members. Thus, it tends to choosing at each step, the merger of the pair of clusters that gives rise to a final cluster that has the smallest diameter. And since the diameter of the merged clusters is minimized at each iteration, the method has a tendency to give rise to compact “globular” (dense and circular or spherical) clusters. This criterion for merging in complete link methods is non-local in the sense that the entire distribution of members in a cluster can influence the choice of the merging. A single element located far away from the majority of the members of one cluster increases significantly the final diameter of two merged clusters which, in turn, may lead to a major change in the final clustering. Thus, the complete-link methods are rather sensitive to outliers, as the single-link methods are sensitive to the noise.

Closely related to the agglomerative linkage methods, the Minimum Spanning Tree (MST) method is an alternate algorithm to build clusters from a given length threshold. When considering a set of spatial points as a complete graph, in which the points are vertices and edges are the euclidean distance between vertices, the MST is the subgraph in which any two
vertices are connected by exactly one simple path (i.e., a connected graph without cycles, by definition a tree) such that the total distance length of its edges is minimal. In this framework, clusters are defined as the remaining connected graphs when deleting the largest lengths of MST above a heuristically chosen threshold (inconsistent edges, Zahn 1971), or the critical length (Gutermuth et al. 2009), such that clusters are then the remaining connected subgraphs (i.e., there is a path between any two points of the subgraph). The same clusters are obtained from the linkage methods agglomerative algorithms employ by cutting the dendrogram at the same critical length, so that each connected component forms a cluster. Indeed, it is well established that the same information required to build the MST of a set of points is contained within the dendrogram generated by the linkage methods (see e.g., Gower & Ross 1969). Therefore, the same drawbacks (high sensitivity to noise and outliers that blur cluster structure) affect both the MST and linkage methods.

C.2. Partitioning clustering

Iterative relocation algorithms use an iterative control strategy to optimize the partition of distribution into clusters. The number of clusters is usually an input parameter for these algorithms, in other words some a priori domain assumption knowledge is required (for a review, see Mirkin 2005). When the number of clusters is given, there are two statistical approaches in order to partition the data in these clusters:

Non-parametric iterative relation algorithms (K-means, K-median) are based on the assumption that clusters correspond to different modes of the probability density, in other words, they are associated with the values that appear more often in a data set. Once an initial partition is given, each cluster is associated with its representative, e.g., its barycenter, centroid, mean point, or by one of the objects of the cluster located near its center. The goal of this class of algorithm is to assign each object to its closest representative and iteratively process to minimize the objective function, as for example the distance of elements to the their closest representatives summed over the over whole set; the iterative process stops when convergence on representatives is reached. For example, the K-means algorithm (Kanungo et al. 2002), a variance-based function, can be shown to minimize within-cluster distance while maximizing between-cluster distance. It should be noted that the assignment of each object to a cluster implies that the induced partition is equivalent to a Voronoi diagram (Inaba et al. 1994; Imai et al. 1995). Thus, the shape of clusters found by a partitioning algorithm is always convex which is a restrictive bias of the method.

On the other hand, parametric iterative relation algorithms (mixture models) are based on the assumption that each cluster is represented by a set of parametric probability density coming from a same family. For example, either a traditional Gaussian distribution, or an “isothermal ellipsoid” (Kuhn et al. 2014), would be employed as a component mixture model method. The “parametric” term refers to the choice of function representing a cluster. Unlike the previous methods, the number of the clusters is required as a given input prior the clustering.

For this class of iterative relocation algorithms, the goal is to estimate (1) the number (and type of the geometrical distribution in the case of parametric methods) of clusters (structure identification), and (2) the parameters of the distributions (weight, variance and mean for each probability distribution associated with a cluster). Broadly speaking, partitioning cluster analysis algorithms are then based on two distinct phases: first is a model fitting phase, whereby the number and the geometrical type of component are a priori chosen, and second is a model validation phase, whereby the set of data are assessed to each cluster according to some cluster validity criterion and to one “optimal” partition hypothesis selected through a maximization technique of an objective function. In most cases, the expectation-Maximization (EM) technique of the (log-)likelihood function is used. The EM technique is able to propose the best local solution in order to smoothly distribute data in each cluster and to constrain the parameters of each probability distribution associated with a cluster. For example, the K-means approach is a special case of EM clustering applied to a mixture of Gaussians when all variances are equal (Gupta & Chen 2011). This is why the K-means method leads to globular shaped clusters.

Iterative relocation methods have several drawbacks: the number of clusters or components have to be set prior to implementation. They are sensitive to first initialization of the partition. A guess of the number of clusters can possibly be inferred based on the Akaike Information Criterion (AIC) such as used in the work of Kuhn et al. (2014). The AIC is a probabilistic indicator that somehow evaluates the gain of introducing some supplementary cluster, and then all the associated free parameters, to improve the model clustering fit evaluated through the global maximum (log-)likelihood function at the expense of model parsimony. Although it is a quantitative criteria, the derived clusters are biased toward convex (i.e., globular) shaped clusters and most importantly, they do not identify arbitrarily sized clusters since they are biased to equipartition. The notion of noisy data is also not considered, in other words, it is a termination requirement that every single object has to be assigned to a cluster.

More sophisticated parametric approaches have been developed to deal with the structure identification based on the optimization of the Bayesian Information Criteria that allows the choice of the “best” model based on direct models comparison, models that are computed simultaneously, each combining different weight of clustering methods (hierarchical and EM partitioning) and various number of components (Fraley & Raftery 1998, 2002, 2007). But even these models do not address the issue of the possible skewness of data distribution in which a single, skewed (or elongated) distribution is described by multiple normal distributions. Moreover, even combining hierarchical and partitional clustering, since there is no explicit notion of noise, the “best” model will be chosen based on complete partition of data even if data are sub-structured, which may be in fact a serious obstacle in determining locally small but significant over-densities.

For these reasons, the iterative relocation partitioning clustering algorithms do not appear adapted to the objective of our study, especially in view of the fact that these algorithms result in a single complete partition of stars into a unique set of disjoint clusters, which may not suit our perception of young star clusters topology that is most of the time complex, highly homogeneous and sub-structured. In summary, given a point-like distribution, the iterative relocation algorithms imposed on by the parametric cluster model, and by the implicit complete partitioning of, prevent the identification of non-globular, irregularly shaped and local small-scale substructures. Hence, partitioning algorithms are ruled out here, and the goal of identifying substructures as local over-densities drives us to the third class of partitioning clustering methods, the density based techniques.

C.3. Density based clustering

Density based algorithms, such as dbscan (Ester et al. 1996), partition a distribution into clusters based on a density
Essentially, they are the formalization of the assumption that clusters are high-density regions surrounded by low-density regions (Hartigan 1975). These algorithms borrow (1) some aspects of the non-parametric and parametric iterative partition cluster methods, and (2) some aspects of MST (single linkage) hierarchical clustering methods. For the former, density clustering algorithms are a non-parametric method, as there is no a priori given function associated with cluster postulating a specific structure for discrete distributions. However, the notion of density probability is not completely ignored since there is an estimation of the local density from a kernel density estimator associated with a given smoothing bandwidth that reduces the complexity. The non-parametric estimation of probability density functions is based on the concept that the value of a density function at a continuity point can be estimated using the sample observations that fall within a small region around that point, known as the Parzen kernel class of density estimates. For a comprehensive review of kernel non-parametric density estimation and the crucial issue of the smoothing parameter (bandwidth) choice see for example Silverman (1986) and Scott (1992). In the case of the dbscan algorithm, a given circular and uniform kernel acts as a given smoothing bandwidth (ε-neighborhood for dbscan) kernel.

For dbscan, clusters are also defined as the maximal connected subgraphs remaining at a cutting scale length threshold, but these subgraphs are composed by selected instances (thus allowing an incomplete partitioning). A significant advantage of using such density based clustering algorithms is that the cluster membership criterion is based on a simple density requirement that allows noisy data to be filtered out, whereas this cannot be done using hierarchical and iterative relocation based partition based algorithms. The dbscan algorithmic implementation appears particularly well suited to our natural perception of star clusters and the study of their structure.